

从基础到实践丛书

SPSS 统计分析

从基础到实践（第2版）

罗应婷 杨钰娟 编著

道然科技 审校

社名:	校次:
责编:	QQ: 826465418
开本:	正文页码:
文前页码:	电话: 59227731
日期:	排版员:
Logo创作室	

電子工業出版社

Publishing House of Electronics Industry

北京 • BEIJING

内 容 简 介

本书是基于 SPSS 17.0 版本进行编写的 SPSS 实用参考手册，全书共 14 章。书中既详细介绍了 SPSS 各菜单的使用方法，又给出了其相应统计方法的基本原理和适用条件。同时，对每个复杂的统计方法，都通过引例讲解说明，这有利于读者学习和真正熟练使用 SPSS 的强大统计功能。

同时，本书最后给出了 SPSS 在各个应用领域的使用实例，其中所用到的统计方法和思想也可以作为读者在处理具体问题时的一个参考。应用实例涵盖了管理决策、生物技术、工程分析、金融系统等领域，所选择的例子不仅具有典型性，而且具有很强的工程参考价值。

本书图文并茂，层次清晰明了，案例丰富多彩，为读者提供了愉快地阅读享受。此外本书光盘还配有精美的 PPT 电子教案，方便教学使用。

本书特别适合希望提升数据统计分析能力的管理者，以及从事统计分析、市场分析、社会学、医药统计分析和金融专业的人员。本书既可以作为利用 SPSS 软件进行数据分析的参考手册，也可以作为各大院校学生学习 SPSS 软件的教材。

未经许可，不得以任何方式复制或抄袭本书之部分或全部内容。
版权所有，侵权必究。

图书在版编目 (CIP) 数据

SPSS 统计分析从基础到实践 / 罗应婷, 杨钰娟编著. —2 版. —北京: 电子工业出版社, 2010.1
(从基础到实践)

ISBN 978-7-121-10010-9

I. S… II. ①罗… ②杨… III. 统计分析—软件包, SPSS IV. C819

中国版本图书馆 CIP 数据核字 (2009) 第 220290 号

责任编辑: 朱沐红

印 刷: 北京智力达印刷有限公司

装 订: 北京中新伟业印刷有限公司

出版发行: 电子工业出版社

北京市海淀区万寿路 173 信箱 邮编 100036

开 本: 787×1092 1/16 印张: 23.75 字数: 542 千字

印 次: 2010 年 1 月第 1 次印刷

印 数: 4000 册 定价: 49.00 元 (含光盘 1 张)

凡所购买电子工业出版社图书有缺损问题, 请向购买书店调换。若书店售缺, 请与本社发行部联系, 联系及邮购电话: (010) 88254888。

质量投诉请发邮件至 zlt@phei.com.cn, 盗版侵权举报请发邮件至 dbqq@phei.com.cn。

服务热线: (010) 88258888。

丛书特色

- 📖 坚持“基础为本源，实践出真知”的写作理念，即首先掌握基本理论和基础技能，然后在实践中锻炼提高。
- 📖 丛书内容“全、精、实用”，只要一本书，初学者就能入门，并完成实际工作。
- 📖 循序渐进地介绍基本知识，通过穿插的小实例，以深入浅出之法加深读者的理解和应用能力，同时强调重点、技巧和注意点。
- 📖 带领读者亲自完成多个项目开发。项目为实践中常用的、典型的应用问题。书中所有实例均调试通过。
- 📖 在配书光盘（或网上下载）中，提供所有练习、实例及实战部分的代码和素材，方便读者学习和使用。
- 📖 贴心顾问技术支持 E-mail：jsj@phei.com.cn，及时解答读者在阅读本书中的问题。



前言

作为全球应用最广泛的统计软件，SPSS 已有 30 余年的历史。它广泛应用于农业、工业、商业、医学、社会学、市场分析、股市行情、旅游业等多个行业和领域。同时，由于 SPSS 的操作简捷、界面美观，受到越来越多的非统计专业人士的青睐。

目前国内也有很多优秀的 SPSS 参考书。但是它们通常要求读者具有较高的统计学基础。对于那些没有系统学习过统计学，但是又迫切需要应用 SPSS 来处理实际问题的读者而言，这些参考书的起点就显得有些高了。因此，本书的编写目的是为各行业中需要应用 SPSS 软件处理实际问题的非统计专业读者提供一本易学、易用的 SPSS 基础教材。

根据作者的经验，非统计专业的读者在使用 SPSS 时通常会遇到如下问题：

- 对 SPSS 软件中提供的各种统计方法的原理和适用条件不清楚，有时胡乱套用方法，出现问题与方法张冠李戴的情况。
- 面对 SPSS 对话框中复杂的选项设置而不知所措。
- 对 SPSS 最后输出的统计图表不知该如何解读。
- 面对有大量数据的复杂问题时不知道如何应用 SPSS 软件来层层剖析数据，解决问题。

本书特色

在本书的编写过程中，作者一直致力于帮助 SPSS 软件的初学者解决以上问题，我们做了以下尝试和努力。

- **结构清晰，易学易用。**本书采用双线索形式安排文章结构。章节设置既符合“输入数据——整理数据——基础统计分析——高级统计分析”这一传统统计分析过程，又严格按照 SPSS 窗口各菜单顺序行文，便于读者尽快熟悉 SPSS 的操作界面。
- **目录明了，便于查询。**由于大部分读者通常只使用 SPSS 的部分功能，因此，本书的目录编排采取了统计名称与功能键名相结合的方法，便于读者快速查询所需方法的操作过程。这样大大克服了一些书籍直接以功能键为目录而使读者无法快速查阅具体所需功能的缺点。
- **详细介绍了 SPSS 的常用功能所对应的对话框。**对其中选项的具体意义、适用情况都有介绍。这样，即使英语水平不高的读者，也可以清晰地了解各个对话框中选项的意义。
- 对于各种统计方法，在具体介绍其在 SPSS 的界面操作之前，都对它们的原理和适

用条件做了详细的介绍。这样，使那些即使没有系统学习过统计学的读者也能够恰当地选择正确的统计方法。

- 对于每种统计方法都给出了具体的例子及其在 SPSS 中的实现，并对最后的输出结果做出了详细解释。初学者可以通过这些例子尽快掌握如何应用 SPSS 来处理实际问题。
- 在本书的最后，给出了多个 SPSS 在行业中的应用实例。这些实例既有编者做过的实际问题，又有其他统计工作者的成功案例。希望读者通过这些案例的学习，能够感受到 SPSS 在解决实际问题中的应用。同时也能学习到其他统计工作者在处理实际问题时的思路，对自己处理实际问题有所启迪。

主要内容

本书以 SPSS 17.0 为基础，但也适用于其他版本的 SPSS 软件。全书共 14 章，分为 SPSS 概述（第 1 章）、数据文件的建立与整理（第 2~5 章）、统计分析（第 6~13 章）和应用实例（第 14 章）四大部分。具体内容包括：SPSS 17.0 概述、SPSS 帮助系统、SPSS 数据文件的建立与编辑、数据整理、统计图、SPSS 报表、描述性统计分析、均值比较与 t 检验、方差分析、相关分析、回归分析、聚类分析、判别分析、因子分析、对应分析、非参数检验及 SPSS 在各领域的应用实例等。

本书既详细介绍了 SPSS 各菜单常用过程的具体使用方法，又给出了其相应统计方法的基本原理和适用条件。同时，对于复杂的统计方法，都通过引例讲解说明，这有利于读者学习和真正熟练使用 SPSS 的强大统计功能。

同时，本书最后给出了 SPSS 在各个应用领域的使用实例，其中所用到的统计方法和思想也可以作为读者在处理具体问题时的一个参考。应用实例涵盖了管理决策、生物技术、工程分析和金融系统等多个领域，所选择的例子具有典型性，而且具有很强的工程参考价值。

本书图文并茂，层次清晰明了，案例丰富多彩，提供给读者愉快的阅读享受。此外本书还配有精美的 PowerPoint 电子教案，方便教师教学使用。

本书的读者对象

本书特别适合希望提升数据统计分析能力的管理者，以及从事统计分析、市场分析、社会学、金融和医药统计分析等专业的人员。既可以作为利用 SPSS 软件进行数据分析的一本参考手册，也可以作为各院校学生学习 SPSS 软件的教材。

光盘使用说明

光盘包含了图书中所有的输入数据文件和配套电子教案，如图 1 所示，并按照图书的章节顺序归类，方便读者查找。

- 配套数据文件的格式包括.sav 和.txt 两种通用格式；
- 包括 8 讲精美的电子教案，文件可以用 Microsoft PowerPoint 软件打开查看；
- 适合 SPSS 10.0 ~ SPSS 17.0 的各个软件版本。

说明：该数据文件是为了提高读者的学习效率而提供的。数据文件的版权属相关出处，仅供学习研究使用。



图 1 光盘中包含的输入数据文件和配套电子教案

致谢与分工

本书由罗应婷、杨钰娟编著，道然科技有限责任公司参与前期的策划和全书的审校工作。北京博文视点资讯有限公司朱沐红女士参与全程的策划工作。参与具体编辑、排版、审校等工作的还有陈军、周维义、刘涛庆、董茜、朱诚、王呼佳、王晓、赵腾化、李佳、刘军华、余松、赵会春等。

所谓厚积薄发，在知识的掌握过程中，我们更加强调循序渐进的学习方法。因此，如果要全面掌握 SPSS 的强大功能，读者需要静下心来，认真阅读本书的各个章节。由于本书编写过程中一直注意各章节间的独立性，因此对那些需要尽快掌握 SPSS 来处理实际问题的读者，只要阅读相关的统计方法部分就可以学会用 SPSS 来处理实际问题的方法。

由于编者水平有限，书中谬误之处在所难免，希望广大读者及时批评指正，如有问题请 Email 联系：sharepub@126.com。

作者
2009 年 9 月



目 录

第 1 篇 SPSS 概述

第 1 章 SPSS Statistics 17.0 基础 2

1.1	SPSS 简介..... 2
1.1.1	SPSS 的产生与发展..... 2
1.1.2	SPSS 17.0 的新特性..... 3
1.1.3	SPSS 与其他常用统计软件比较..... 3
1.1.4	SPSS 的主要应用领域简介..... 4
1.2	SPSS 17.0 窗口简介..... 4
1.2.1	数据编辑窗口 (SPSS Statistics Data Editor) 4
1.2.2	结果浏览窗口 (SPSS Statistics Viewer) 7
1.2.3	程序编辑窗口 (SPSS Statistics Syntax Editor) 10
1.2.4	VBs 宏程序编辑窗口 Script 10
1.3	SPSS 17.0 的帮助系统..... 11
1.3.1	对话框上的 Help 按钮..... 11
1.3.2	主题词获得帮助——Topics 过程..... 11
1.3.3	新手入门——Tutorial 过程..... 12
1.3.4	实例学习——Case Studies 过程..... 13
1.3.5	统计教练——Statistics Coach 过程..... 13
1.3.6	语法指南——Command Syntax Reference 过程..... 14

1.3.7	算法介绍——Algorithms 过程..... 14
-------	-----------------------------

1.3.8	访问 SPSS 官方主页..... 15
-------	----------------------

1.4	本章小结..... 15
-----	--------------

第 2 篇 数据文件的建立与整理

第 2 章 SPSS 数据文件的建立与编辑..... 18

2.1	变量定义与数据输入..... 18
2.1.1	定义新变量..... 18
2.1.2	数据的录入与编辑..... 22
2.2	数据文件的创建与保存——File 菜单详解..... 22
2.2.1	新建 SPSS 数据文件..... 22
2.2.2	导入其他类型数据文件..... 22
2.2.3	保存数据文件..... 25
2.2.4	File 菜单的其他命令..... 26
2.3	数据文件的编辑与管理——Edit/Utilities 菜单详解..... 27
2.3.1	Edit 菜单详解..... 27
2.3.2	Utilities 菜单详解..... 29
2.4	本章小结..... 31

第3章 SPSS 数据文件的整理..... 32

- 3.1 数据文件整理概述.....32
 - 3.1.1 数据文件的整理在实际工作中的重要性.....32
 - 3.1.2 一个数据文件整理的案例.....32
- 3.2 数据文件的整理——Data 菜单详解.....33
 - 3.2.1 观测量排序——Sort Case 过程.....33
 - 3.2.2 数据文件转置——Transpose 过程.....34
 - 3.2.3 数据格式重排——Restructure 过程.....35
 - 3.2.4 数据文件合并——Merge File 子菜单.....37
 - 3.2.5 数据分类汇总——Aggregate 过程.....41
 - 3.2.6 数据文件的拆分——Split File 过程.....44
 - 3.2.7 选择观测量——Select Cases 过程.....46
 - 3.2.8 观测量加权——Weight Cases 过程.....48
 - 3.2.9 Data 菜单其他过程简介.....49
- 3.3 变量的变换和计算——Transform 菜单详解.....49
 - 3.3.1 变量计算——Compute Variable 过程.....49
 - 3.3.2 变量值标识——Count Values within Cases 过程.....52
 - 3.3.3 变量重新赋值——Recode into Same Variables/ Recode Into Different Variables 过程.....54
 - 3.3.4 变量值秩排序——Rank Cases 过程.....57
 - 3.3.5 Transform 菜单其他过程简介.....60
- 3.4 本章小结.....60

第4章 SPSS 统计图形..... 61

- 4.1 统计图形概述..... 61
 - 4.1.1 Graphs 菜单简介..... 61
 - 4.1.2 常用统计图形简介..... 65
- 4.2 常见统计图形..... 66
 - 4.2.1 条形图 (Bar Charts)..... 66
 - 4.2.2 线图 (Line Charts)..... 73
 - 4.2.3 面积图 (Area Charts)..... 75
 - 4.2.4 饼图 (Pie Charts)..... 75
 - 4.2.5 高低图 (High-Low Charts)..... 76
 - 4.2.6 帕累托图 (Pareto Charts)..... 77
 - 4.2.7 质量控制图 (Control Charts)..... 79
 - 4.2.8 箱图 (Boxplot) 与误差条图 (Error Bar)..... 80
 - 4.2.9 金字塔图 (Population Pyramid)..... 81
 - 4.2.10 散点图 (Scatter/Dot)..... 83
 - 4.2.11 直方图 (Histogram)..... 83
 - 4.2.12 P-P 图和 Q-Q 图..... 85
 - 4.2.13 ROC 曲线..... 87
 - 4.2.14 时间序列图 (Time Series Charts)..... 89
- 4.3 SPSS 图形编辑..... 93
 - 4.3.1 图形编辑概述..... 93
 - 4.3.2 图形基本设定——Edit 菜单..... 94
 - 4.3.3 图形高级设定——Options 菜单和 Elements 菜单..... 95
- 4.4 交互式统计图形..... 97
 - 4.4.1 交互式统计图形概述..... 97
 - 4.4.2 交互式条图的界面..... 97
 - 4.4.3 交互式条图实例..... 99
- 4.5 本章小结..... 100

第5章 SPSS 报表..... 101

- 5.1 简单记录报表——Reports 子菜单..... 101
 - 5.1.1 在线分析处理——OLAP 过程..... 101

5.1.2	观测量汇总——Case Summaries 过程	105
5.1.3	生成商务报表——Report Summaries in Rows/Columns 过程	108
5.2	高级报表——Tables 子菜单	115
5.2.1	定义复选变量集——Multiple Response Sets 过程	115
5.2.2	定制报表——Custom Tables 过程	117
5.3	本章小结	122

第3篇 统计分析

第6章 描述性统计分析 124

6.1	描述性统计量	124
6.1.1	描述性统计量	124
6.1.2	Descriptive Statistics 子菜单 概述	125
6.2	频数分布表分析——Frequencies 过程	126
6.2.1	Frequencies 过程的操作界面	126
6.2.2	引例	128
6.3	最基础的统计量分析—— Descriptive 过程	130
6.3.1	Descriptive 过程的操作界面	130
6.3.2	引例及结果解释	131
6.4	探索性分析——Explore 过程	131
6.4.1	Explore 过程的操作界面	132
6.4.2	引例及结果解释	133
6.5	列联表分析——Crosstabs 过程	139
6.5.1	Crosstabs 过程的操作界面	139
6.5.2	引例	142
6.5.3	结果解释	143
6.6	相对比描述——Ratio 过程	144
6.6.1	Ratio 过程的操作界面	144

6.6.2	引例及结果解释	146
6.7	本章小结	148

第7章 均值比较与 t 检验 149

7.1	t 检验简介	149
7.1.1	t 检验的概念及一般步骤	149
7.1.2	t 检验的类型	149
7.2	均值描述——Means 过程	150
7.2.1	Means 过程的操作界面	150
7.2.2	引例及结果解释	152
7.2.3	分组变量的层次说明	153
7.3	单样本 t 检验—— One-Sample T Test 过程	154
7.3.1	单样本 t 检验的一般步骤	154
7.3.2	One-Sample T Test 过程的 操作界面	155
7.3.3	引例及结果解释	155
7.4	独立两样本 t 检验 ——Independent-Sample T Test 过程	156
7.4.1	独立两样本 t 检验的一般步骤	157
7.4.2	Independent-Sample T Test 过程的操作界面	157
7.4.3	引例及结果解释	159
7.5	配对样本 t 检验—— Paired-Sample T Test 过程	160
7.5.1	配对样本 t 检验一般步骤	160
7.5.2	Paired-Sample T Test 过程的 操作界面	161
7.5.3	引例及结果解释	162
7.6	本章小结	163

第8章 方差分析 164

8.1	方差分析简介	164
8.1.1	方差分析的提出	164
8.1.2	方差分析的基本概念	164

8.1.3	方差分析的类型	165
8.2	单因素方差分析—— One-Way ANOVA 过程	166
8.2.1	单因素方差分析简介	166
8.2.2	One-Way ANOVA 过程的 操作界面	167
8.2.3	引例及结果解释	169
8.3	多因素方差分析—— Univariate 过程 (1)	172
8.3.1	多因素方差分析简介	172
8.3.2	Univariate 过程的操作界面	175
8.3.3	引例及结果解释	180
8.4	协方差分析—— Univariate 过程 (2)	183
8.4.1	协方差分析简介	183
8.4.2	引例及结果解释	184
8.4.3	小结	189
8.5	本章小结	189

第 9 章 相关分析 190

9.1	相关分析简介	190
9.1.1	相关分析的概念	190
9.1.2	Correlate 子菜单概述	191
9.2	两变量相关分析——Bivariate 过程	191
9.2.1	两变量相关分析简介	191
9.2.2	Bivariate 过程的操作界面	193
9.2.3	引例及结果解释	194
9.3	偏相关分析——Partial 过程	197
9.3.1	偏相关分析简介	197
9.3.2	Partial 过程的操作界面	198
9.3.3	引例及结果解释	199
9.4	距离分析——Distances 过程	201
9.4.1	距离分析简介	201
9.4.2	Distances 过程的操作界面	201
9.4.3	引例及结果解释	205
9.5	本章小结	206

第 10 章 回归分析 207

10.1	回归分析简介	207
10.1.1	回归分析的概念	207
10.1.2	回归分析的应用	208
10.1.3	回归分析的类型	208
10.1.4	回归分析的一般步骤	209
10.2	线性回归——Linear 过程	210
10.2.1	线性回归简介	210
10.2.2	Linear 过程的操作界面	212
10.2.3	一元线性回归的例子	217
10.2.4	多元线性回归的例子	220
10.2.5	小结	224
10.3	曲线拟合——Curve Estimation 过程	225
10.3.1	曲线拟合简介	225
10.3.2	Curve Estimation 过程的操作 界面	225
10.3.3	引例及结果解释	227
10.4	二分类变量 Logistic 回归—— Binary Logistic 过程	230
10.4.1	Logistic 回归简介	230
10.4.2	Binary Logistic 过程的操作 界面	231
10.4.3	引例及结果解释	234
10.4.4	小结	238
10.5	非线性回归——Nonlinear 过程	239
10.5.1	非线性回归简介	239
10.5.2	Nonlinear 过程的操作界面	239
10.5.3	引例及结果解释	243
10.5.4	小结	246
10.6	本章小结	246

第 11 章 聚类分析与判别分析 248

11.1	聚类分析与判别分析相关原理 简介	248
------	---------------------	-----

11.1.1	聚类分析	248
11.1.2	判别分析	248
11.2	K-均值聚类分析——K-means Cluster 过程	249
11.2.1	K-均值聚类法基本原理	249
11.2.2	K-means Cluster 过程界面 操作介绍	249
11.2.3	引例及结果解释	252
11.3	系统聚类法——Hierarchical Cluster 过程	254
11.3.1	系统聚类法基本原理	254
11.3.2	Hierarchical Cluster 过程界面 操作介绍	254
11.3.3	引例及结果解释	257
11.4	两步聚类法——TwoStep Cluster 过程	263
11.4.1	两步聚类法基本原理	263
11.4.2	TwoStep Cluster 过程界面 操作介绍	264
11.4.3	引例及结果解释	266
11.5	判别分析——Discriminant 过程	272
11.5.1	判别分析基本原理	272
11.5.2	Discriminant 过程界面 操作介绍	273
11.5.3	引例及结果解释	276
11.6	本章小结	280

第 12 章 因子分析与对应分析 281

12.1	因子分析——Factor Analysis 过程	281
12.1.1	因子分析基本原理	281
12.1.2	Factor Analysis 过程界面 操作介绍	283
12.1.3	引例及结果解释	286
12.2	简单对应分析——Correspondence Analysis 过程	296
12.2.1	简单对应分析基本原理	296

12.2.2	Correspondence Analysis 过程 界面操作介绍	297
12.2.3	引例及结果分析	299
12.3	最优尺度分析——Optimal Scaling 过程初步认识	301
12.4	本章小结	303

第 13 章 非参数检验 304

13.1	非参数检验相关原理简介	304
13.1.1	非参数检验的概念	304
13.1.2	非参数检验的优缺点	305
13.1.3	非参数检验的类型	305
13.2	分布类型的检验	306
13.2.1	卡方检验——Chi-Square 过程	306
13.2.2	二项分布检验——Binomial 过程	314
13.2.3	游程检验——Runs 过程	316
13.2.4	单个样本的 K-S 检验——1-Sample K-S 过程	319
13.3	分布位置检验	322
13.3.1	两个独立样本分布位置检验——2 Independent Samples 过程	322
13.3.2	多个独立样本分布位置检验——K Independent Samples 过程	325
13.3.3	两个相关样本分布位置检验——2 Relate Samples 过程	328
13.3.4	多个相关样本分布位置检验——K Relate Samples 过程	331
13.4	本章小结	334

第 4 篇 应用实例

第 14 章 SPSS 在各领域的应用实例 336

14.1	SPSS 在房地产决策中的应用	336
------	-----------------	-----

14.1.1	问题描述	337	14.3.1	问题描述	350
14.1.2	问题建模	337	14.3.2	问题建模	351
14.1.3	模型的验证	344	14.3.3	模型的检验	356
14.2	SPSS 在生物模型中的应用	344	14.4	SPSS 在证券分析中的应用	357
14.2.1	问题描述	345	14.4.1	问题描述	357
14.2.2	问题建模	345	14.4.2	问题建模	358
14.2.3	模型的讨论	349	14.4.3	模型的讨论	366
14.3	SPSS 在工程问题中的应用	350			

PART

第 1 篇 SPSS 概述

- 第 1 章 SPSS Statistics 17.0 基础

第 1 章 SPSS Statistics 17.0 基础

作为全书的开篇，本章详细介绍了 SPSS 的基础知识、常用窗口以及帮助系统的用法。当然，对 SPSS 的初学者来说，这些基础理论知识是必要的，却也是枯燥的。因此，建议将这一章作为后面章节的辅助材料来阅读。本章的内容涵盖如下几方面：

- SPSS 简介；
- SPSS 17.0 窗口简介；
- SPSS 17.0 的帮助系统。

1.1 SPSS简介

在国际学术界有一条不成文的规定：在国际学术交流中，凡是用 SPSS 软件完成的计算和统计分析，可以不用说明算法。仅从这点就足以说明在国际上使用 SPSS 的广泛程度。

1.1.1 SPSS的产生与发展

SPSS 原意是 Statistical Package for the Social Sciences，意为社会科学统计软件包。但是最近，伴随着 SPSS 产品服务领域的扩大和服务深度的增加，SPSS 公司已决定将其英文全称更改为 Statistical Product and Service Solutions，即统计产品与服务解决方案。

SPSS 统计软件系统最早是由美国斯坦福大学的三位学生于 1968 年开发的。基于这一系统，1975 年他们在芝加哥成立了 SPSS 公司总部，推出了 SPSS 中小型机版本 SPSSX。1984 年，SPSS 总部终于推出了世界上第一个统计分析软件微机版本 SPSS/PC，显然，这个微机版本极大地扩展了它的应用范围。20 世纪 90 年代，SPSS 又推出了 Windows 版本，从 SPSS 5.0 开始，一直到现在的 SPSS 17.0，它的功能一直在不断增强，以满足各种客户的不同需求，世界上许多有影响的报刊杂志纷纷就 SPSS 的自动统计绘图、数据的深入分析、使用方便、功能齐全等方面给予了高度的评价与赞赏。

SPSS Statistics 是一个组合式软件包，它集数据整理、分析功能于一身。它的基本功能包括数据管理、统计分析、图表分析、输出管理等等。SPSS 还有专门的绘图系统，该系统功能非常强大，可以根据数据绘制各种图形。SPSS 17.0 输出的图形美观、大方，如图 1-1 所示，这是用 SPSS 17.0 所绘制的一个 3D 饼图。

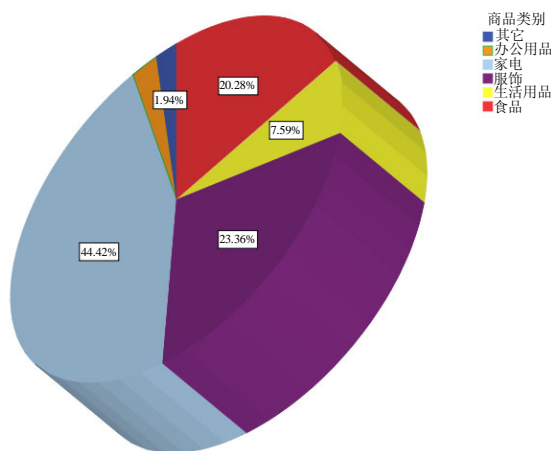


图 1-1 3D 饼图

1.1.2 SPSS 17.0 的新特性

迄今为止，SPSS 公司的最新产品就是 SPSS 17.0，它在开放性、集成性和满足企业级用户需求等方面展示了超群的性能。作为企业级软件，SPSS 17.0 可以更有效地处理各企业用户的海量数据。它新增加的主要功能如下。

- 可切换的界面语言：除了能够更改以前版本中提供的输出语言外，SPSS 17.0 还更改用户界面语言，方便不同国籍的用户操作和使用。
- 更优秀的图表功能：用户可以采用图形板直观地表示模板创建的图形和图表，方便用户选择个性化图形。
- 更简便的数据处理：新增了多重插补功能修补缺失值，引入了中位数函数，增强了分类汇总功能。
- 更强大的统计分析功能：新增了最邻近元素分析、RFM 分析，同时增强了分类回归分析的功能。
- 更多输出导出格式选项和更多导出内容控制。

1.1.3 SPSS与其他常用统计软件比较

除了 SPSS 以外，常用的统计软件还包括 SAS、S-Plus、Stata 和 Eviews 等等，这些统计软件凭借它们各自的优点，在统计分析领域中发挥着重要的作用。

SAS：被誉为国际上的标准统计软件和最权威的组合式统计软件。由于它是为专业统计分析人员设计的，因而它具有功能强大，应用灵活多样的特点，为许多专业人士所喜爱，但对于非专业人士来说，它的人机对话界面不太友好，学习起来也比较困难。

S-Plus：作为 S 语言的后续发展，它在应用上以理论研究和统计建模为主。它的优点是具有强大的统计功能和绘图能力。当然，同样因为它的专业性，使用这款软件需要有较好的数理统计背景，并且对编程能力要求极高。

Stata: 这款软件与其他款软件相比，较为小巧，它的统计分析能力很强，绘图也很美观，但是它不提供对话框界面，而是使用命令行方式操作。

Eviews: 这款软件的主要贡献是在计量经济学上，特别是在时间序列分析和面板数据分析上有优势。

SPSS: 最突出的特点就是它使用 Windows 的窗口方式展示各种管理和分析数据的功能，使用对话框展示各种功能选择项。它可以直接读取 Excel 和 DBF 数据文件，而且现在已经推广到多种操作系统上。它的操作界面非常友好，输出的结果清晰、直观。整个系统易学易用，只要对统计分析原理有基本的了解就可以使用，因而是非专业统计人员的首选统计软件。

1.1.4 SPSS的主要应用领域简介

迄今 SPSS 软件已有 40 余年的历史。全球约有 25 万家产品用户，它们广泛分布于农业、工业、商业、医学、交通运输、公检法、社会学、市场分析、股市行情、军事地理和旅游业等多个领域和行业，是世界上应用最广泛的专业统计软件之一。可以这么说，有需要数据分析的地方，就可以用到 SPSS。

1.2 SPSS 17.0 窗口简介

本节将介绍 SPSS 17.0 中几类主要的窗口，其中最常用的窗口就是数据编辑窗口和结果浏览窗口。

1.2.1 数据编辑窗口（SPSS Statistics Data Editor）

启动 SPSS 17.0 后，首先弹出的窗口就是数据编辑窗口，如图 1-2 所示，它是由标题栏、菜单栏、工具栏、数据编辑窗口和状态栏构成，下面来依次介绍。

1. 标题栏

显示 SPSS 当前所打开的数据文件名，若没有数据文件被打开，则显示空数据编辑窗口，如图 1-2 所示，标题为“Untitled1 [DataSet 0] - SPSS Data Editor”（未命名 1 [数据集 0] - SPSS 统计数据编辑器），其中的数据集概念是自 SPSS 14.0 后引入的。SPSS 17.0 允许同时打开多个数据编辑窗口，这些数据编辑窗口可以显示不同的数据文件，也可以显示同一个数据文件的多个数据集。在多个打开的数据编辑窗口中，只有当前活动的数据编辑窗口是用标题栏最左边有一个绿色的加号表示。

2. 菜单栏

标题栏下面紧接着的就是菜单栏，如图 1-3 所示，上面显示了所有的一级菜单。如果要进行某项具体的操作，需要单击相应的一级菜单，然后在弹出的子菜单中选择命令选项，下面来介绍这些一级菜单的功能。

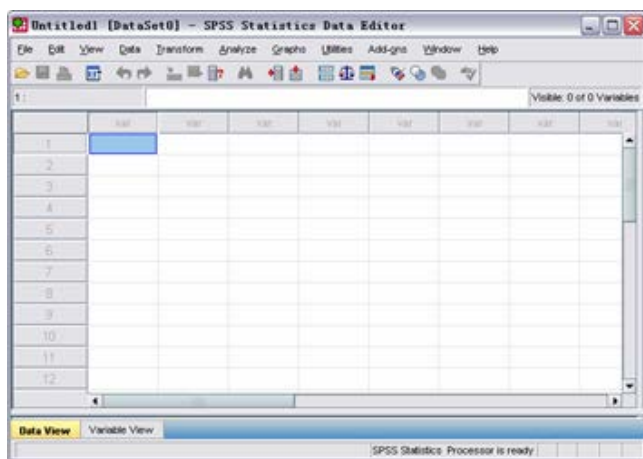


图 1-2 数据编辑窗口

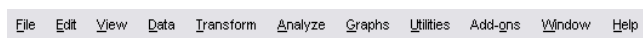


图 1-3 数据编辑窗口的菜单栏

- **【File】**（文件）

文件管理菜单。负责新建文件、读入文件或数据库、保存文件、标记文件为只读文件、重命名数据库、显示数据文件信息、打印等操作。其中特别指出两个功能：一个是**【Cache Data】**（建立数据缓冲区），它可以将数据载入内存，大大提高运行的速度；另一个是**【Recently Used File】**（最近使用过的其他文件）和**【Recently Used Data】**（最近使用过的数据文件），将两者分别放置，为用户提供了便捷打开常用文件的方式。

- **【Edit】**（编辑）

编辑菜单。对文件数据进行选择、复制、粘贴、删除、查找，还可以插入新变量、新观测量，以及利用**【Options】**（选项）菜单修改整个 SPSS 设置等操作。

- **【View】**（显示）

显示菜单。进行窗口外观控制、自定义工具栏、显示字体设置、显示或隐藏格子、显示变量标签、在数据浏览和变量浏览之间切换等操作。

- **【Data】**（数据）

数据整理菜单。进行数据变量的定义、复制数据或数据集、定位观测量、分类观测量、转换、重构变量、合并其他文件数据等操作。

- **【Transform】**（转换）

变量整理菜单。进行数值计算、重新编码、缺失值替代、创建时间序列、产生随机数等操作。

- **【Analyze】**（分析）

统计分析菜单。应用各种统计方法对当前窗口中的数据进行分析。

- **【Graphs】**（图表）

图表菜单。根据当前数据绘制和编辑各种统计图表。

- **【Utilities】**（实用选项）

实用选项菜单。进行变量列表、控制输出管理系统、输出文件信息、定义和使用变量集合、运行菜单编辑器等操作。

- **【Add-ons】**（增加模块）

增加模块菜单。包括应用、服务、编程扩充和统计指导，该菜单提供了强大的高级统计分析和数据挖掘功能。

- **【Window】**（窗口）

窗口管理菜单，进行窗口拆分、最小化、切换窗口等操作。

- **【Help】**（帮助）

帮助菜单，提供 SPSS 系统帮助、在线指南、统计分析指导等功能。

3. 工具栏

紧接在菜单栏下面的就是数据编辑窗口的工具栏，工具栏中的按钮都能在菜单中找到相应的命令，如图 1-4 所示。这是 SPSS 数据编辑窗口的默认工具栏，这些按钮从左到右依次表示：打开文件、保存文件、打印、再次调出对话框、向后撤销、向前撤销、定位到观测量、定位到变量、显示变量、查找数据、插入观测量、插入变量、拆分文件、加权观测量、选择观测量、变量值标签、使用变量集合、显示所有变量以及拼写检查。



图 1-4 数据编辑窗口的工具栏

SPSS 工具栏中默认的按钮都是最常用的按钮。当然，根据不同的需要，用户也可以自定义工具栏。具体的方法是，先执行 **【View】 / 【Toolbars】 / 【Customize】** 命令，这时就会弹出如图 1-5 所示的 **【Show Toolbars】**（显示工具栏）对话框。然后单击 **【Edit】** 按钮，弹出如图 1-6 所示的 **【Edit Toolbar】**（定制工具栏）对话框。



图 1-5 显示工具栏对话框

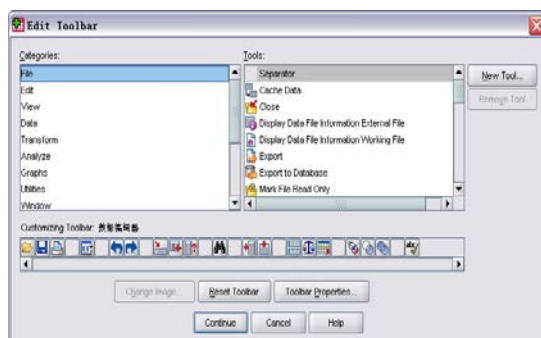



图 1-6 定制工具栏对话框


这时就可以根据自己的需要清除不需要的按钮，添加需要的按钮。比如，添加一个清除按钮 ，具体操作如下：

执行【View】/【Toolbars】/【Customize】命令，弹出【Show Toolbars】对话框
单击【Edit】按钮，弹出【Edit Toolbar】对话框

【Edit Toolbar】对话框：


Categories 列表框组：


选中“Edit”菜单，Tools 列表框中显示出所有属于 Edit 的命令及其图标

单击选项“ Delete”，拖曳至下方的 Customizing Toolbar 对话框。

单击【Continue】按钮，【Edit Toolbar】对话框定义完成

单击【OK】按钮，定义完成

执行以上操作之后，则成功添加清除按钮  至工具栏。

若想在工具栏中删除打印按钮 ，则执行如下操作：

执行【View】/【Toolbars】/【Customize】命令，弹出【Show Toolbars】对话框
单击【Edit】按钮，弹出【Edit Toolbar】对话框

【Edit Toolbar】对话框：

将删除按钮  拖曳出 Customizing Toolbar 框

单击【Continue】按钮【Edit Toolbar】对话框定义完成

单击【OK】按钮定义完成

这时，工具栏中就已删除了打印按钮 。

4. 数据编辑窗口

在工具栏的下方是数据编辑窗口，可以在其中进行数据的编辑操作。

5. 状态栏

数据编辑窗口下方就是状态栏，显示 SPSS 程序此时的工作状态，如图 1-2 所示，此时 SPSS 程序的状态是“SPSS Processor is ready”，即 SPSS 程序已经做好准备。

1.2.2 结果浏览窗口（SPSS Statistics Viewer）

SPSS 的一个显著特点就是其输出结果简洁易读，便于编辑。如图 1-7 所示，这就是一个 SPSS Statistics Viewer（结果浏览窗口）。它的组成部分也包括了标题栏、菜单栏、工具栏、输出结果编辑窗口和状态栏五部分。与数据编辑窗口相比，它的菜单栏和工具栏有一些不同之处，这是它们功能上的差异造成的。

如图 1-7 所示，SPSS Statistics Viewer 的输出结果编辑窗口被分成了左右两个部分。左边部分称为 Outline view（结构视图或大纲视图），右边部分显示详细的统计结果，包括统计图、统计表和文本输出结果等。左右两边的元素是一一对应的。可以说左侧 Outline view 是右侧具体输出结果的目录。

下面来具体介绍统计结果浏览窗口的菜单栏和工具栏。因篇幅有限，这里就不再一一叙述每个菜单的具体用处，而是着重介绍它区别于数据编辑窗口的几个重要功能。

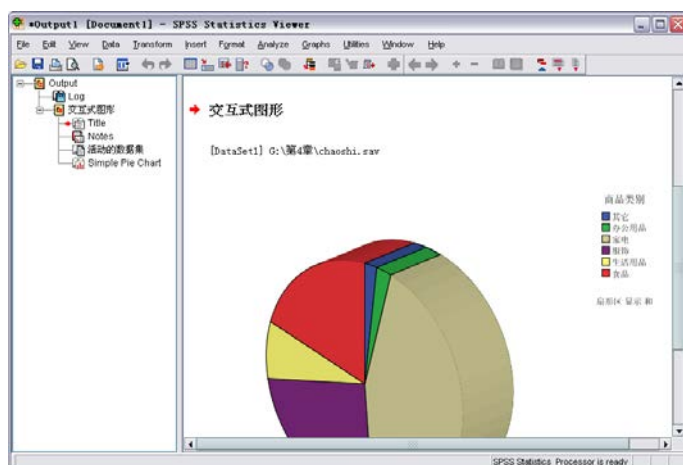


图 1-7 结果浏览窗口

1. 【File】菜单

• 【Export】（导出结果）

【File】菜单下的子菜单，在结果浏览窗口中有重要的作用，可以方便用户将结果发送到网上、打印以及在其他软件中进行再编辑。执行【File】/【Export】命令，弹出如图 1-8 所示的【Export Output】对话框，在这里可以选择导出的具体内容、导出结果的文件类型及设置、导出的位置等详细信息。

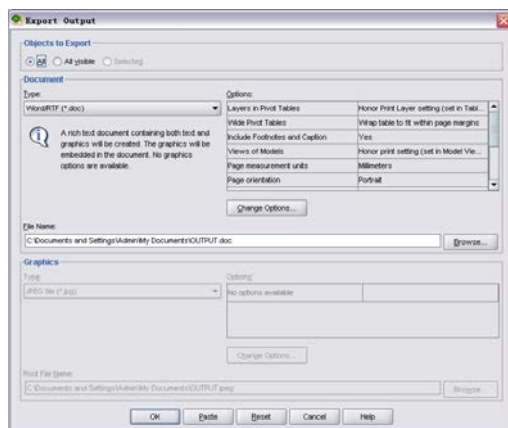


图 1-8 【Export Output】对话框

前面已经提过，SPSS 17.0 在文件导出方面的功能更加强大，主要包括以下几点。

- 通过【Change Option】按钮设置各类导出文件的用户自定义样式。
- 导出的 Word 文档可换行或缩小宽表。
- 可在 Excel 工作簿中创建新工作表或向现有工作表追加数据。

这些改进有效地改善了以前 SPSS 某些结果导出文件不太规范的缺点，更有利于用户操作。

2. 【Edit】菜单

• 【Paste Special】（特殊粘贴）

【Edit】菜单下的子菜单。若要复制结果浏览窗口中的图形，执行【Edit】/【Paste Special】命令，弹出如图 1-9 所示的【Paste Special】对话框。此时对话框中有两个选项：SPSS Statistics Viewer Object 和 image。若选择 SPSS Statistics Viewer Object，则把图形粘贴成可编辑的对象格式。若选择 image，则把图形粘贴为纯图片格式。若要复制结果浏览窗口中的表格或文字，则执行【Edit】/【Paste Special】命令，弹出如图 1-10 所示的【Paste Special】对话框。该对话框中有两个选项：SPSS Statistics Viewer Object 和 Text，分别表示将选中的图表粘贴为可编辑的对象格式和纯文本格式。

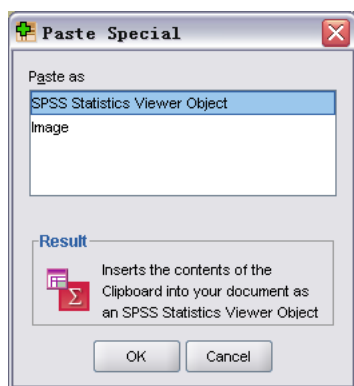


图 1-9 【Paste Special】对话框

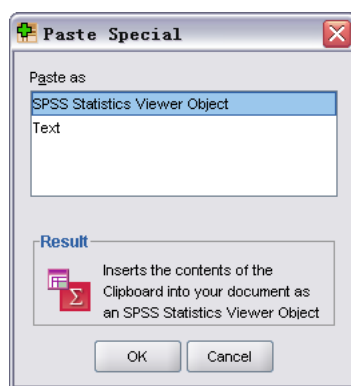


图 1-10 【Paste Special】对话框

3. 【Insert】（插入）

插入菜单，这是结果浏览窗口所特有的，可以在窗口中红箭头所指位置设置页码断点、清除页码断点（用于打印时分页）、插入标题、名称、文本内容、二维图表、三维图表、或任意对象等。

4. 【Format】（格式）

格式菜单，这也是结果浏览窗口特有的，上面有三个选项，分别是将选中对象左对齐、居中、右对齐，这是为打印效果所设置的。

接着介绍结果浏览窗口的默认工具栏中每个按钮的意义。如图 1-11 所示，这些按钮从左到右的功能依次是：打开结果窗口、保存当前结果窗口、打印、打印预览、导出结果、最近使用对话框、向后撤销、向前撤销、定位到数据、定位到观测量、定位到变量、变量信息、使用变量集合、所有变量、选择最后一个输出结果、关联脚本、创建编辑脚本、运行脚本、指定输出窗口（当有多个结果浏览窗口同时打开时需要选定一个作为输出窗口）、提升所选内容为上级目录、降低所选内容为下级目录、展开所选目录、折叠所选目录、显示所选内容、隐藏所选内容、插入大标题、插入名称、插入文本内容。



图 1-11 结果浏览窗口的工具栏

1.2.3 程序编辑窗口（SPSS Statistics Syntax Editor）

SPSS 中的任何操作过程都可以转化为相应的程序语句在 SPSS Syntax Editor（程序编辑窗口）中输出。

具体转化是通过各个 SPSS 过程的对话框中的【Paste】（粘贴）按钮来实现的。例如，在数据文件已打开的情况下，执行【Analyze】/【Compare Means】/【Means】命令，弹出如图 1-12 所示的对话框，选好变量后，单击图中箭头所指的【Paste】按钮，相应的程序就会出现在如图 1-13 所示的程序编辑窗口中。在程序编辑窗口，执行【Run】/【All】命令，就等价于执行如图 1-12 所示的对话框中的操作。

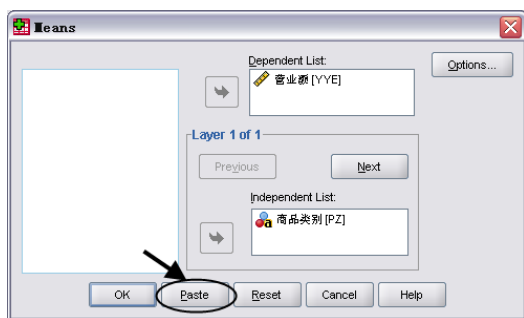


图 1-12 【Means】对话框

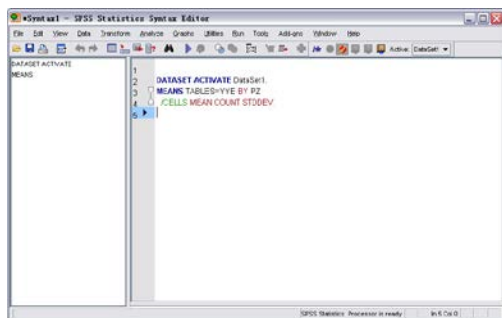


图 1-13 程序编辑窗口

若用户对 SPSS 非常熟悉，还可以通过执行【File】/【New】/【Syntax】命令，自己新建一个程序编辑窗口，然后输入命令语句来完成某些自定义的功能。

1.2.4 VBs宏程序编辑窗口Script

在 SPSS 数据编辑窗口或者结果浏览窗口中执行【File】/【Open】/【Script】命令，然后在【Open File】（打开文件）对话框中打开 Scripts 文件夹中任意一个以“sbs”为扩展名的文件，就出现如图 1-14 所示的一个宏程序编辑窗口。用户还可以通过执行【File】/【New】/【Script】命令新建一个 Script 文件，然后自行编程生成脚本。

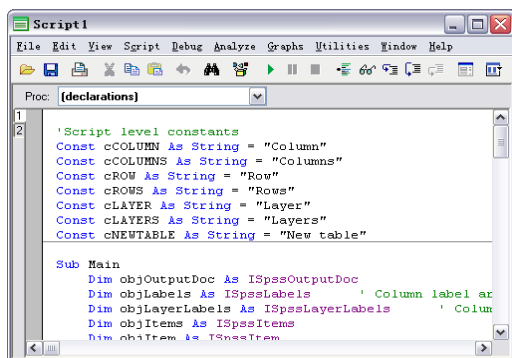


图 1-14 宏编辑窗口

1.3 SPSS 17.0 的帮助系统

SPSS 提供了详尽而丰富的在线帮助系统，任何一本介绍 SPSS 的书都没有它详尽。了解如何获得帮助，对学习和使用 SPSS 非常重要。

1.3.1 对话框上的 Help 按钮

几乎每个 SPSS 对话框上面都有一个【Help】按钮，单击它就会弹出【Help】对话框，上面详细介绍了这个对话框上的各个选项、框组的作用。以结果浏览窗口的【Export】子菜单为例，在如图 1-8 所示的【Export Output】对话框中，单击【Help】按钮，弹出如图 1-15 所示的对话框，上面详细介绍了【Export Output】对话框上各个选项的含义。



图 1-15 【Help】按钮帮助

1.3.2 主题词获得帮助——Topics 过程

单击数据编辑窗口或者结果浏览窗口的【Help】菜单就会弹出如图 1-16 所示的下拉菜单，下面将对它从上到下依次讲解。

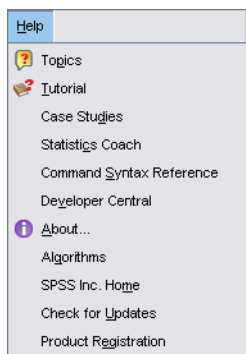


图 1-16 数据编辑窗口中的 Help 菜单

单击【Topics】菜单后，弹出如图 1-17 所示的对话框，这就是主题词帮助内容，你可以通过在索引选项卡中输入关键字来直接获得相关帮助内容，也可以在目录中查找你感兴趣的课题。

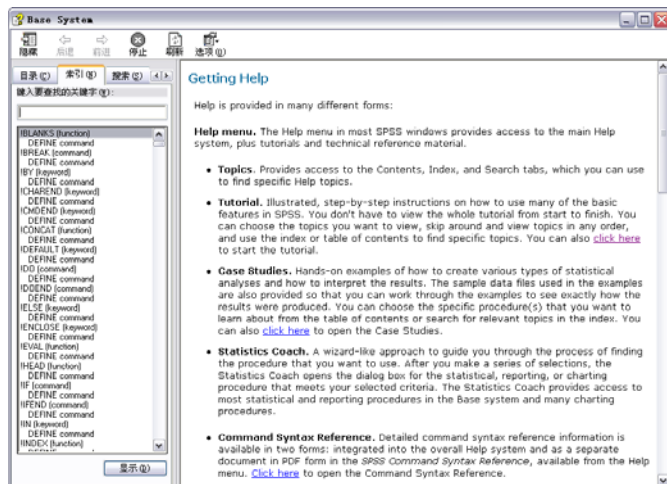


图 1-17 主题词帮助

1.3.3 新手入门——Tutorial 过程

Tutorial 的中文意思就是指南，它以动画的形式解释 SPSS 中的各种操作，非常适合初学者。执行【Help】/【Tutorial】命令后，就会出现指南对话框的主页，上面列出了所有的操作主题，单击某项主题，弹出如图 1-18 所示的图文并茂的对话框，对话框顶部显示帮助的主题，左侧是具体的操作视图，右侧是具体操作步骤，对话框右下方的四个按钮的意义从左到右分别是：搜索、回到目录界面、向前翻页和向后翻页。

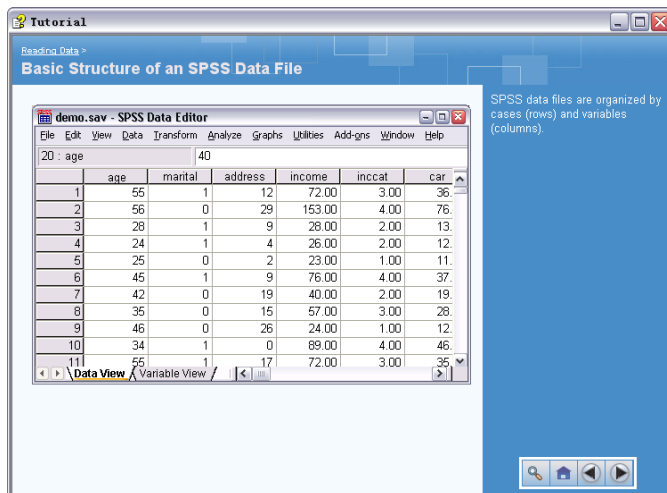


图 1-18 指南帮助

1.3.4 实例学习——Case Studies过程

最快的学习方法莫过于通过实例进行具体操作了，【Case Studies】就是通过向用户展示具体的实例操作过程来帮助用户尽快掌握 SPSS 的功能。执行【Help】/【Case Studies】命令，进入实例学习对话框的主页，单击某项具体操作选项，弹出如图 1-19 所示对话框，它的结构与指南对话框非常相似。

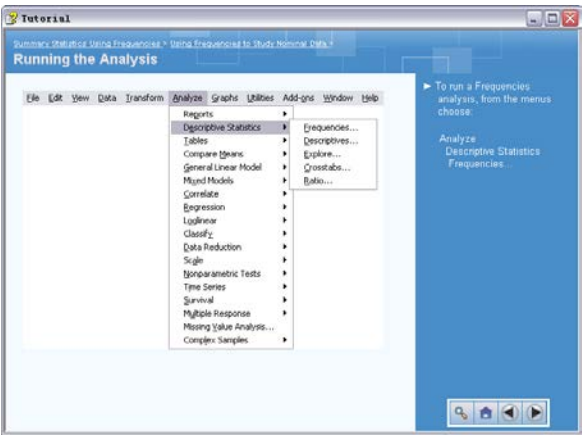


图 1-19 实例学习

1.3.5 统计教练——Statistics Coach过程

如果用户对 SPSS 还不熟悉，却要马上运用它实现某些功能，这时【Statistics Coach】（统计教练）就起大作用了。执行【Help】/【Statistics Coach】命令，弹出如图 1-20 所示对话框，上面将通过一个个的对话框详细询问你想要让 SPSS 实现什么功能，选择完毕后，桌面上弹出两个对话框，一个是教练的建议，也就是如何实现这个方法，另一个对话框就是相应功能实现过程的对话框，你就可以直接在里面操作了。

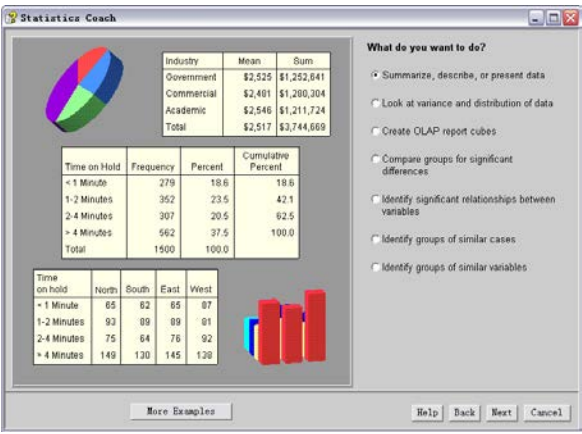


图 1-20 统计教练

1.3.6 语法指南——Command Syntax Reference过程

执行【Help】/【Command Syntax Reference】命令，弹出一个 pdf 文档，如图 1-21 所示，这就是 SPSS 的所有语法指南。

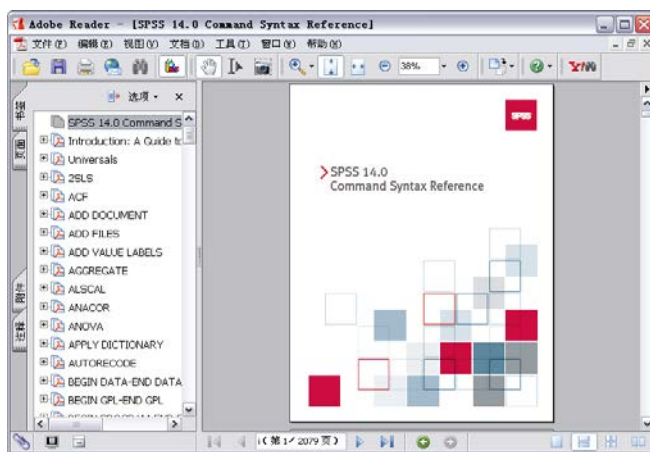


图 1-21 相关命令语法

1.3.7 算法介绍——Algorithms过程

执行【Help】/【Algorithms】命令，弹出如图 1-22 所示的文件夹，里面有关于每个算法的 pdf 文档，其中包括该算法的基本原理、计算方法和参考资料。

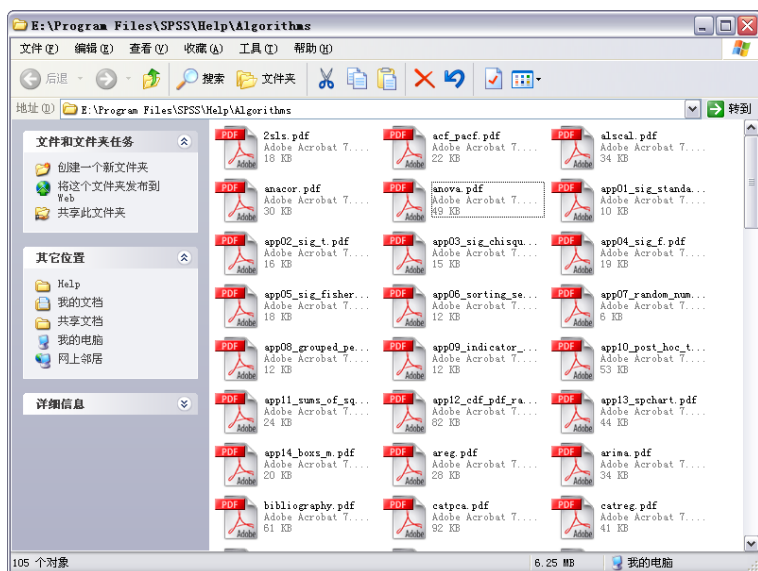


图 1-22 算法文件夹

1.3.8 访问SPSS官方主页

确定你能上互联网后，执行【Help】/【SPSS Inc. Home】命令，IE 就会自动进入 SPSS 官方主页，你就可以了解更多关于 SPSS 的信息了

1.4 本章小结

作为全书第一章，本章简单介绍了 SPSS 的基础知识。通过本章的学习，读者需要掌握以下知识：

- SPSS 数据编辑窗口和结果浏览窗口；
- SPSS 帮助系统的基本操作。

同时，对 SPSS 发展历史、与常用统计软件的比较以及应用领域能有一个初步了解。

PART

第 2 篇 数据文件的建立与整理

- 第 2 章 SPSS 数据文件的建立与编辑
- 第 3 章 SPSS 数据文件的整理
- 第 4 章 SPSS 统计图形
- 第 5 章 SPSS 报表

第 2 章 SPSS 数据文件的建立与编辑

上一章对 SPSS 的基础知识、常用窗口和帮助系统做了详细介绍。在用 SPSS 分析数据处理实际问题的时候，第一步就是建立数据文件。因此本章首先介绍 SPSS 数据文件的建立与编辑，本章的内容涵盖以下几大方面：

- 变量定义与数据输入；
- 数据文件的创建与保存——File 菜单详解；
- 数据文件的编辑与管理——Edit / Utilities 菜单详解。

2.1 变量定义与数据输入

本节主要介绍新变量的定义以及数据的输入与编辑。这些主要是针对用新数据文件定义原始数据变量而言的。而由原始变量生成新变量主要是由【Transform】菜单完成，将在下一章介绍。

2.1.1 定义新变量

打开 SPSS 之后，直接进入数据编辑窗口。SPSS 的数据编辑窗口分为 Data View（数据）和 Variable View（变量）两个视图子窗口。用左键单击左下方的 Variable View 标签就可以得到如图 2-1 所示的变量窗口。

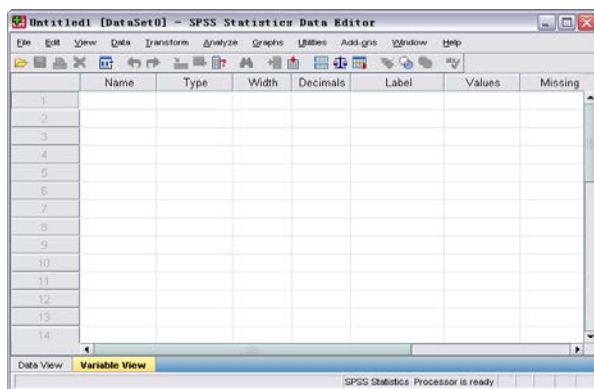


图 2-1 SPSS 变量定义窗口

SPSS 变量窗口主要用来定义变量的名称、类型、标签等。窗口中每一行代表一个变量，每一列（栏）代表变量所对应的一个属性。现在具体讲解变量窗口各列的用法。

1. Name 栏

输入变量名。SPSS 17.0 的变量长度可多达 64 位。但是由于老版本的 SPSS 变量名长度应在 8 位之内，为了避免与老版本及其他软件出现兼容问题，变量名一般仍控制在 8 位以内且尽量避免采用中文。而必要的中文说明可以放在 Label 栏。需要注意的是变量名不能与 SPSS 的保留字相同。SPSS 的保留字包括：all、by、eq、ge、gt、le、lt、ne、not、or、to、with。同时，系统不区分变量名的大小写。

2. Type 栏

选择变量类型。左键单击 Type 栏后边  按钮，则会弹出如图 2-2 所示的对话框。

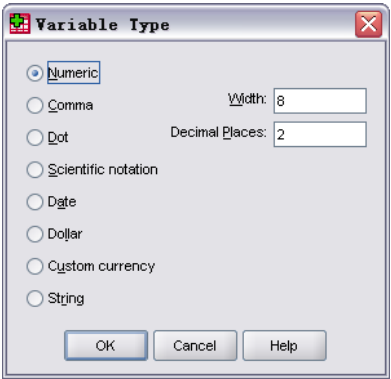


图 2-2 【Type】对话框

SPSS 最基本的变量类型有数值型、字符型和日期型三种。根据细化条件，Type 对话框共提供了八种可选数据类型。各种数据类型的说明如表 2-1 所示。

表 2-1 SPSS 变量类型说明

英文名	中文名	说明
Numeric	标准数值型变量	默认总长度 8 位，小数位 2 位
Comma	带逗号的数值型变量	默认总长度 8 位，小数位 2 位。其值在显示时整数部分从右至左每三位用一个逗号作分隔符
Dot	带圆号的数值型变量	默认总长度 8 位，小数位 2 位。其值在显示时整数部分从右至左每三位用一个圆点作分隔符
Scientific notation	科学记数法数值型变量	默认总长度 8 位，小数位 2 位。变量值可以有指数部分也可以没有。指数部分用 E 或 D 表示且可带“+”、“-”号
Date	日期型变量	既可表示日期又可表示时间，用户可根据实际情况自行选择
Dollar	带美元符号的数值型变量	其值在显示时前面可带“\$”
Custom currency	自定义数值型变量	用户可以自定义变量类型，但是此项一般不用
String	字符型变量	默认总长度为 8 位

在不指定变量类型的情况下，当数据文件输入为数值时，变量类型默认为 **Numeric**。当输入为字符时，变量类型默认为 **String**。

3. Width 栏

设置变量宽度。一般无需调整，直接采取默认值。它的大小可通过 **Width** 栏后边的微调按钮调整，也可以通过图 2-2 中的 **【Width】** 选项调整。

4. Decimals 栏

设置小数位。默认为两位，其大小的调整方法同 **Width**。


5. Label 栏

定义变量名标签，即对变量名给出具体的解释和说明。考虑到与老版本的兼容问题，SPSS 17.0 变量名最好依然限制为 8 位以内且尽量避免中文，但这样有时就不能完全描述清楚变量的信息。遇到这种情况的时候，**Label** 栏就大有用武之地了。利用 **Label** 栏，不仅可以对变量详细说明，而且还可以采用中文说明，大大方便了用户对变量的理解。

6. Values 栏

定义变量值标签，即对特定变量输入值给出其详细说明。这在大量的数据输入中是十分有用的。

比如有来自三个国家的多名游客，要输入如图 2-3 所示的一组游客信息。最常用的方法是直接输入游客姓名和其对应的国籍。下边的操作给出如何利用 **Values** 栏，简化输入过程。

STEP 01 在 **Variable View** 窗口下，左键单击变量“country”的 **Values** 栏后边的  按钮，弹出如图 2-4 所示的 **【Values】** 对话框。在 **【Values】** 对话框中输入如图 2-4 所示文字。其中“**Value**”表示变量值，“**Label**”表示变量标签，通过“**Add**”将变量值与变量标签的对应关系选入下边白框内。单击 **【OK】** 按钮，完成变量值标签的定义，其中：变量值“1”表“America”，“2”表“England”，“3”表“Australia”。

	Name	Country	var
1	Mary	American	
2	Mike	Australia	
3	Jane	Australia	
4	Tracy	England	
5	David	American	

图 2-3 Values 栏数功能演示 1

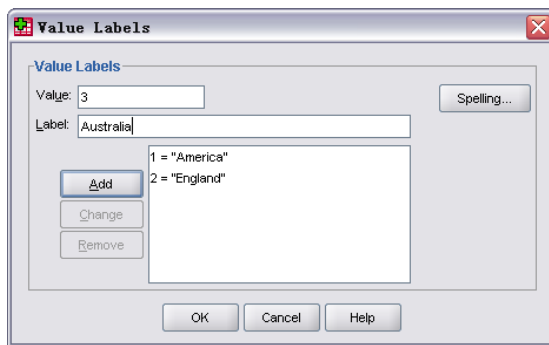


图 2-4 Values 对话框

STEP 02 切换回 **Data View** 窗口，按如图 2-5 所示方式输入数据。

STEP 03 执行【View】/【Value Labels】命令，图 2-5 中的数据就会自动变成如图 2-6 所示数据。

	name	country	var
1	Mary	1	
2	Mike	3	
3	Jane	3	
4	Tracy	2	
5	David	1	

图 2-5 Values 栏数功能演示 2

	name	country	var
1	Mary	America	
2	Mike	Australia	
3	Jane	Australia	
4	Tracy	England	
5	David	America	

图 2-6 Values 栏数功能演示 3

通过比较图 2-5 与图 2-6 可以发现，使用 Values 可以简化数据输入过程。对于频繁出现且复杂的数据，就可以通过定义变量值标签简化输入过程，再通过执行【View】/【Value Labels】命令来还原其值。

7. Missing 栏

定义缺失值。SPSS 缺失值默认为“.”。左键单击 Missing 栏，弹出如图 2-7 所示对话框。第一项表示无缺失值；第二项表示不连续缺失值，至多可自定义三个；第三项表示缺失值范围并可给定一个缺失值。

8. Columns 栏

定义列宽。

9. Align 栏

定义变量值显示方式，默认为左对齐。



图 2-7 【Missing】对话框

10. Measure 栏

定义变量的测量尺度。准确定义测量尺度，才能为数据分析和交互式绘图做好准备。SPSS 的测量尺度主要有标度测量、有序测量和名义测量三种，具体如表 2-2 所示。

表 2-2 SPSS 的测量尺度

名 称	符 号	说 明	对应的变量类型	举 例
标度测量 (Scale)		测量水平最高，包括的信息最多。测量值之间既可做减法运算，又可做除法运算来比较大小	只能是数值型	学生分数、人的身高的具体数值
有序测量 (Ordinal)		信息量低于标度测量，只能保存测量值之间的一种有序关系	可以是数值型，也可以是字符型	人的身高等级：“高”、“中”、“矮”。虽然不能如标度测量一样做减法或除法运算，但是 对于身高有“高”>“中”>“矮”
名义测量 (Nominal)		测量水平最低，取值仅代表一定的分类或标识，测量值之间没有大小可言	可以是数值型，也可以是字符型	人的性别分为“男”、“女”。无法比较究竟是哪个优于哪个

以上详细介绍了变量定义窗口各栏的作用。在实际问题中最常用的为 Name、Type、Label、Values 和 Measure。其余一般可以使用默认值或用其他方式替代其功能。

2.1.2 数据的录入与编辑

SPSS 数据录入是所有统计软件中最方便的一个。左键单击 SPSS 变量视图窗口的 Data View 标签就可以切换到数据录入窗口，然后直接通过键盘录入数据。数据的行列转换功能也可直接由键盘的“上”、“下”、“左”、“右”键完成。相信用户在亲自使用了 SPSS 之后就会对其便捷的数据录入方式有深刻的印象。

除了直接录入数据之外，SPSS 还可以直接复制粘贴 Excel 和 Word 表格中的数据。同时 SPSS 中的数据也可以直接粘贴到 Excel 和 Word 之中。这大大方便了用户制作文本。

SPSS 数据录入有一项特殊功能就是连续粘贴相同值。譬如，有需要连续录入多个相同变量值的时候，可以先录入一项，然后单击鼠标右键，在弹出的快捷菜单中选择【copy】命令。再拖动鼠标选中所有要录入该值的单元格，单击鼠标右键，在弹出菜单中选择【paste】命令。这时就可以发现所有的单元格都已经同时粘贴上该值，而无需一一粘贴了。

2.2 数据文件的创建与保存——File菜单详解

上一节介绍了 SPSS 中变量定义和数据录入。数据只有形成了一个数据文件之后才能进行统计分析。因此，本节首先介绍了数据文件的创建。

观察如图 2-8 所示的【File】菜单可以发现，SPSS 共有 4 种创建数据文件的方法：新建数据文件、直接打开已有数据文件、使用数据库查询方式打开数据文件、从文本导入数据文件。本节将一一介绍这四种方法。

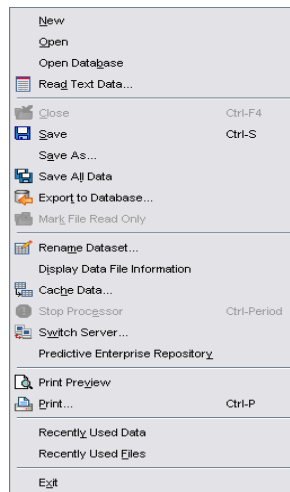


图 2-8 【File】菜单

2.2.1 新建SPSS数据文件

启动 SPSS 之后就自动打开了一个后缀为“*.sav”的空数据文件。按照 2.1 节的方法定义变量和输入数据就可以建立一个通常的 SPSS 数据文件了。

在已打开数据文件的情况下要新建数据文件，执行【File】/【New】/【Data】命令，即可创建一个新的空数据文件。这都是非常简单的操作，此处跳过。但是值得一提的是，自 SPSS 14.0 开始，SPSS 可以在一个 SPSS 进程中同时打开多个数据文件，这比以往的一个进程只能打开一个数据文件是一大进步，更加方便用户多窗口切换操作。

2.2.2 导入其他类型数据文件

从其他数据文件导入数据主要有直接打开、用数据库查询方式打开和从文本文件导入三种方法。这三种方法中最简单的是直接打开。但有的数据文件不能直接打开，这时就可以采用数据库查询的方式打开。而从文本文件导入数据则是一种针对纯文本数据文件的打开方式。

1. 直接打开

SPSS 可直接打开很多类型的数据文件，执行【File】/【Open】/【Others】命令，弹出【Open File】对话框，左键单击“文件类型”，即可看到 SPSS 所能打开的数据文件类型，如表 2-3 所示。

表 2-3 SPSS 能直接打开的数据文件类型

文件 名	描 述	备 注
SPSS (*.sav)	新版本 SPSS 数据文件	SPSS17.0 默认格式
SPSS/PC+ (*.sys)	老版本 SPSS 数据文件	--
Systat (*.syd)	*.syd 格式的 Systat 数据文件	--
Systat (*.sys)	*.sys 格式的 Systat 数据文件	--
SPSS Portable (*.por)	SPSS 的 ASCII 数据文件	通常不用 SPSS 来打开
Excel (*.xls)	Excel 数据文件	常用，Excel 下文件仅当保存为“Excel 4.0 工作表”才能打开
Lotus (*.w*)	Lotus 数据文件	--
SYLK (*.slk)	SYLK 数据文件	--
dBase (*.dbf)	dBase 数据文件	Foxpro 下的 dbf 文件需转换为 dBase (*.dbf)才能打开
SAS Long File Name (*.sas7bdat)	SAS 长数据文件	--
SAS Short File Name (*.sd7)	SAS 短数据文件	--
SAS v6 for Windows (*.sd2)	SAS 老版本数据文件	--
SAS v6 for UNIX (*.ssd01)	SAS 新版本数据文件	--
SAS transport (*.xpt)	SAS 便携文件	--
Text (*.txt)	文本数据文件	--
Data (*.dat)	文本数据文件	--

用户选中自己想要打开的文件类型和文件名，左键双击该文件可以打开该文件。

2. Open Database打开

执行【File】/【Open Database】/【New Query】命令，弹出如图 2-9 所示的【Database Wizard】对话框。这里显示了所有可以打开的数据源类型。

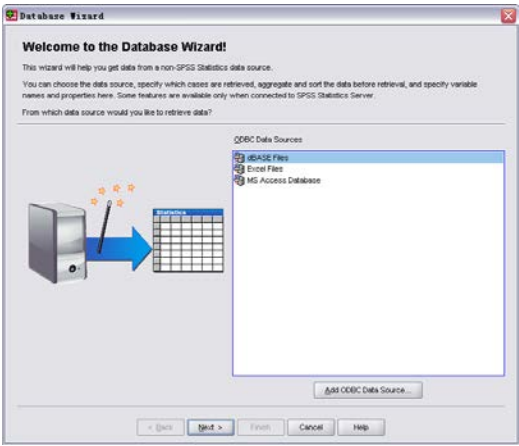


图 2-9 【Database Wizard】对话框

用户根据打开文件的向导选择要打开的文件类型并逐步打开文件。其实从前面的讲解可以发现，直接打开方式已经可以打开很多常见类型的数据文件了。但是当与 SQL Server、DB2、Oracle 等大型数据库进行数据交换时，直接打开数据文件往往是做不到的。所以此时就要使用数据库查询的方式打开数据文件。所以用数据库查询的方式打开文件可以说是一种“没有办法时的办法”。

3. Read Text Data打开文本文件

执行【File】/【Read Text Data】命令与执行【File】/【Open】/【Others】命令再选择“*.txt”的效果一样，都可以跳出相同的【Open File】对话框，如图 2-10 所示。在 SPSS 中将文本文件的打开单独列出来其实只是为了和老版本的界面兼容。由于打开文本文件是经常会遇到的一种特殊情况，所以这里详细讲解一下其具体操作过程。

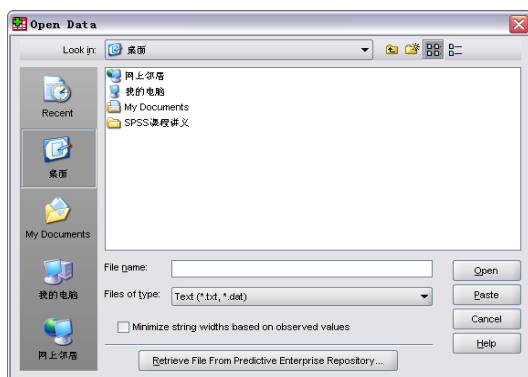


图 2-10 【Open File】对话框

选定一个文件后进入一个分为 6 步的打开向导。

STEP 01 如图 2-11 所示，主要是预览数据文件的读入格式，此时通常情况是与预定义格式不一致的。所以系统默认选择“No”，并单击【Next】按钮。

STEP 02 如图 2-12 所示，第二步主要用来定义变量名属性，有两个单选选项：首先选择变量名的排列方式（用特定字符分割“Delimited”，还是固定宽度“Fixed Width”）；然后选择变量名是否包含在文件之中。

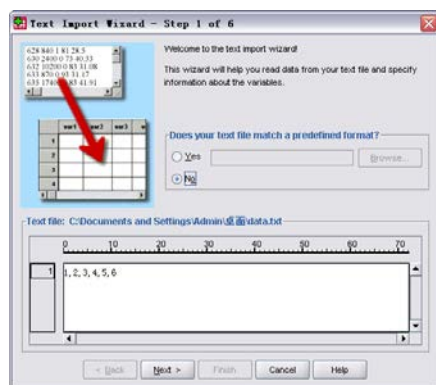


图 2-11 文本文件打开第一步

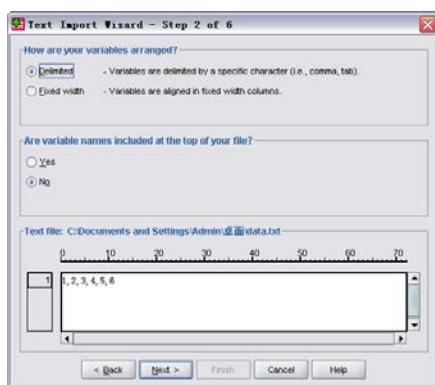


图 2-12 文本文件打开第二步

STEP 03 若在 Step 2 中, 变量名排列方式选择为“Delimited”, 则弹出如图 2-13 所示对话框, 该对话框用来定义导入记录的属性。从上至下依次定义记录开始行数、记录排列方式(一行代表一条记录还是几个固定数目的值代表一条记录)以及要导入的记录数。

STEP 04 如图 2-14 所示, 主要用于选择数据的分割符号和数据采用何种文本限定符号。此时, 系统已经自动识别并选择了打开文本的对应选项。因此, 通常情况下用户直接单击下一步而无需再做调整。

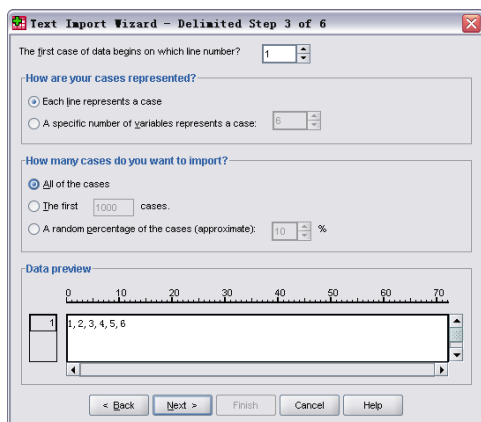


图 2-13 文本文件打开第三步

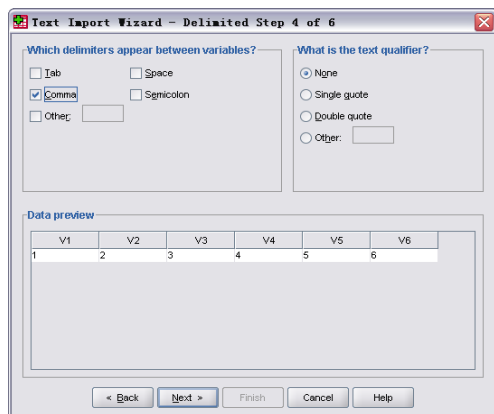


图 2-14 文本文件打开第四步

STEP 05 如图 2-15 所示, 为各个变量命名并定义变量类型。

STEP 06 询问数据文件的保存, 通常直接选择默认情况, 如图 2-16 所示。最后单击【Finish】按钮, 则成功导入了一个“*.txt”的文本文件。

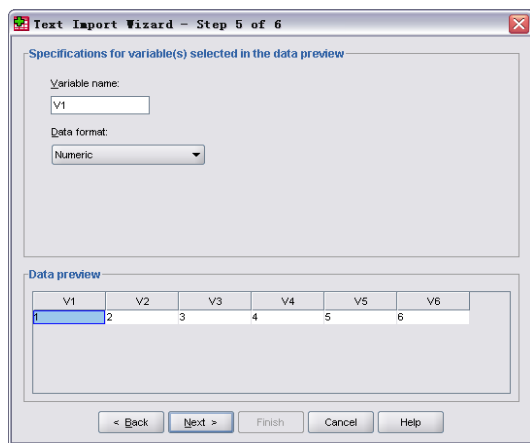


图 2-15 文本文件打开第五步

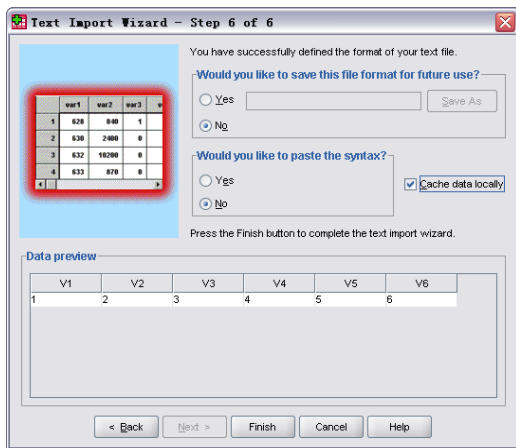


图 2-16 文件打开第六步

2.2.3 保存数据文件

打开数据文件之后, 如果做了相应的修改那么对文件的保存是必不可少的。SPSS 同其他常用软件一样, 对于数据文件有两种保存方式, 即【File】/【Save】或【File】/【Save As】。

但是 SPSS 有一项特殊的选择性保存功能。所谓选择性保存，即只保存数据文件中的部分变量。在用 SPSS 做数据整理和统计分析的时候，会生成一些中间变量。有时用户只希望保存原始数据，有时用户则希望保存新生成的变量，此时就可以使用选择性保存功能。

执行【File】/【Save All Data】命令，弹出【Save Data As】对话框。左键单击该对话框下的【Variables】按钮，弹出如图 2-17 所示保存变量选择对话框。

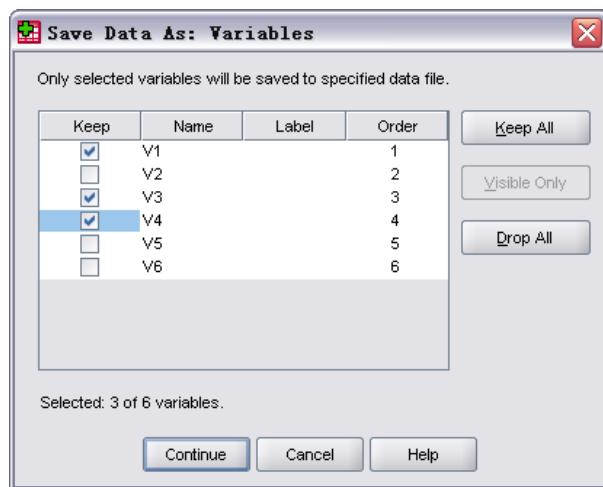


图 2-17 【Save Data As: Variables】对话框

系统默认为保存所有变量。左键单击【Drop All】按钮则任何变量都不保存。一般情况是左键单击变量名前的 ☒ 来决定是否保存该变量。☒ 表示保存，☐ 表示不保存。这种方法可使用户自由剔除对最终结果不重要的中间变量，大大节约了存储空间。

2.2.4 File菜单的其他命令

前面介绍了【File】菜单创建、打开和保存数据文件这几项主要功能。表 2-4 列出了【File】菜单的其他功能。

表 2-4 File 菜单其他命令简介

名 称	作 用
Export to Database	将数据导出到数据库
Mark File Read only	将数据文件设为只读状态，对于不能修改的重要数据文件可做此项设置。左键单击该项之后该选项就变成【Mark File Read Write】，再单击【Mark File Read Write】则将数据文件设置还原为读写状态
Rename Dataset	为数据文件重命名
Display Data File Information	显示数据文件信息。包括【Working File】和【External File】。选择前者则显示打开数据文件变量的详细信息。选择后者则可显示其他外部 SPSS 文件的变量详细信息
Cache Data	建立数据缓冲区，主要针对远程调用大型 SQL 数据库的情况才会起到比较明显的作用
Stop Processor	停止当前 SPSS 命令，主要用于当处理大型数据出现系统忙而无法跳出命令的时候。但是该功能并不是对所有命令都有用，比如计算变量就无法停止

续表

名 称	作 用
Switch Sever	主要用于联网的计算机处理子机之间的切换，一般单机版用户不会用到此项功能
Predictive Enterprise Repository	预测企业库，一般单机版用户也不会用到此项功能
Print&Print View	打印及打印预览
Recently Used Data	显示最近使用的数据文件
Recently Used Files	显示最近使用的其他类型文件

2.3 数据文件的编辑与管理——Edit/Utilities菜单详解

上一节介绍了数据文件的创建与保存。本节主要介绍数据文件的编辑与管理。其中【Edit】菜单同一般 Office 软件的功能近似。而【Utilities】菜单则是 SPSS 很有特色的一项数据编辑和管理功能。

2.3.1 Edit菜单详解

SPSS 通过如图 2-18 所示的【Edit】菜单实现基本的数据和变量编辑功能。

如图 2-18 所示，【Edit】菜单可分为以下 4 个部分。

1. 基本操作

主要包括向后撤销操作【Undo】、向前撤销操作【Redo】、数据剪贴【Cut】、数据复制【Copy】、数据粘贴【Paste】和数据清除【Clear】，这些功能都和一般的 Excel 操作一致。这里就不在赘述。

需要一提的是 SPSS 有一项新增功能：变量粘贴【Paste Variables】。这项功能主要用来复制粘贴变量。当左键单击数据编辑窗口左下方的 Variable View 标签时，数据编辑窗口切换到变量视图窗口。如图 2-19 所示，整行选中某一个变量，单击鼠标右键，在弹出的快捷菜单中选择【copy】命令。

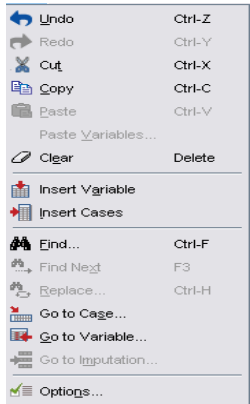


图 2-18 【Edit】菜单

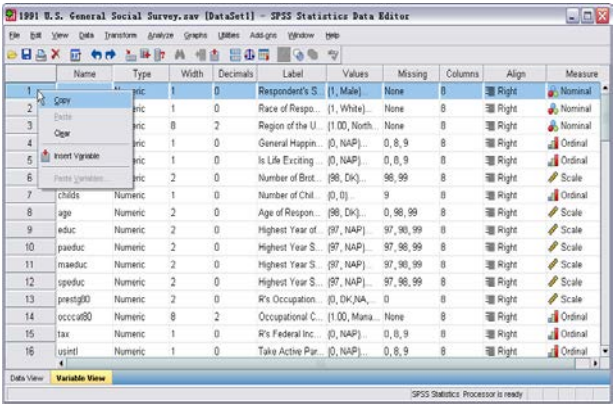


图 2-19 变量定义窗口

再选中变量定义窗口中的任意一行，执行【Edit】/【Paste Variables】命令，弹出如图 2-20 所示对话框。

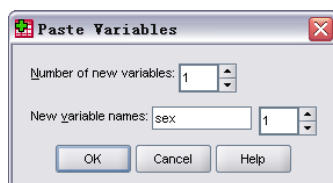


图 2-20 【Paste Variables】对话框

【Paste Variables】有两个选项，其中 Number of new variables 栏用来确定将要复制粘贴的变量个数；New variable names 栏用来确定粘贴生成的新变量的变量名。单击【OK】按钮，则生成了相应个数的新变量。此时，新变量的所有属性都和原来复制变量的属性相同。

2. 插入

【Insert Variables】用来在当前鼠标所在区域左边插入一列新变量。【Insert Case】用来在当前鼠标所在区域上边插入一行新的记录。

3. 查询

包括查找【Find】、查找下一个【Find Next】、查找并替换【Replace】、定位到某一行数据【Go to Case】以及定位到某一列变量【Go to Variable】。

4. 选项

执行【Edit】/【Options】命令，弹出如图 2-21 所示的【Options】对话框，该对话框主要用来设定 SPSS 软件的基本设置。

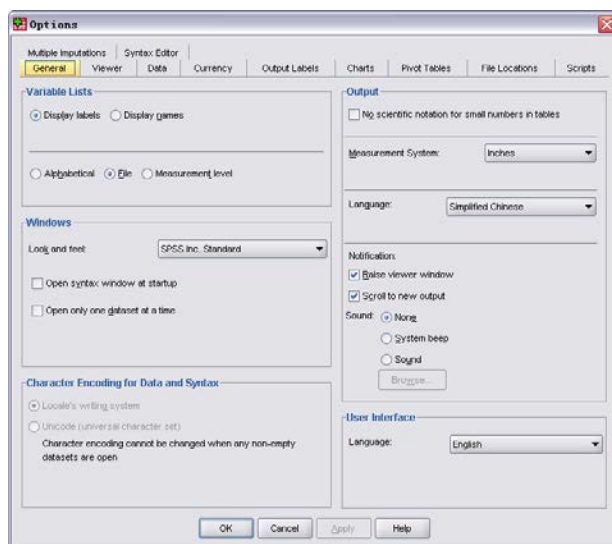


图 2-21 【Options】对话框

SPSS 的【Edit】菜单实现的都是一些比较常规的编辑功能，而其真正便捷的数据编辑与管理功能都是由下一小节所介绍的【Utilities】菜单实现的。

2.3.2 Utilities菜单详解

【Utilities】菜单有着独特的数据编辑与管理功能。现在对 SPSS 17.0【Utilities】菜单的几个常用选项一一介绍。

1. 【Variables】

用于显示变量信息。对一个打开的数据文件，执行【Utilities】/【Variables】命令，弹出如图 2-22 所示的【Variables】对话框。

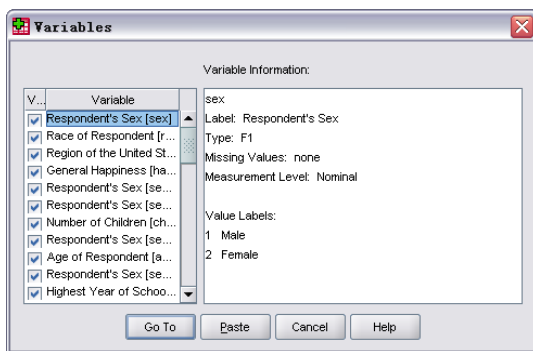


图 2-22 【Variables】对话框

图 2-22 右侧的【Variable Information】框显示的是左侧选中变量的具体信息。本图中显示的是变量“sex”的信息。其中变量标签空缺为“Respondent's sex”。变量类型为 F1，表示长度为 1 位且无小数位的数值型。若变量类型长度为 1 位且有 2 位小数的数值型，则用 F1.2 表示。若变量类型为字符型则用 A 表示，为日期型用 DATA 表示。测量尺度为 Nominal。变量值标签为“1”表示“Male”，为“2”表示“Female”。左键单击【Go To】按钮，系统鼠标作用区域直接跳到变量“sex”所在列。左键单击【Paste】按钮则将变量“sex”复制到变量编辑窗口。

2. 【OMS control Panel】

对于 SPSS 的使用者来说，如果操作过程中涉及一些中间数据表的话，就可以使用 SPSS 的 OMS (Output Management System) 功能。OMS 是自 SPSS 12.0 开始的新增功能。该功能可以将中间生成的数据表自动存到指定类型的文件中去。这些文件包括“*.sav”、“*.XML”、“HTML”和“Text”4 类。但由于属于高级功能，所以目前只能通过交互式编程实现。具体操作可以参考 SPSS 的帮助文件。

3. 【OMS Identifiers】

执行【Utilities】/【OMS Identifiers】命令，弹出如图 2-23 所示对话框。在使用 OMS 功能进行编程的时候，可以通过粘贴该对话框的命令模块到 Syntax 窗口来简化编程过程。

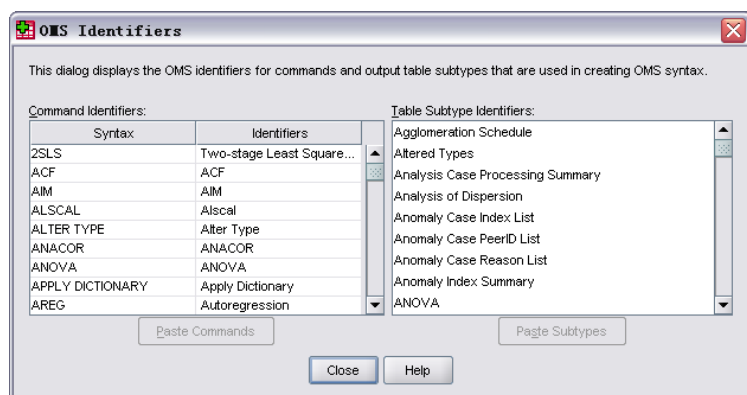


图 2-23 OMS 定义窗口

4. 【Data File Comments】

主要用于用户自己添加一些数据文件的信息，如图 2-24 所示。选中“Display Comments in output”，则添加的结果就会在 output 窗口中显示。用户每进行一次新的添加后，系统都会自动在添加内容的后边生成一个添加日期的标识信息。

5. 【Define Variable Sets】

定义变量集。执行【Utilities】/【Define Variable Sets】命令，弹出如图 2-25 所示对话框。

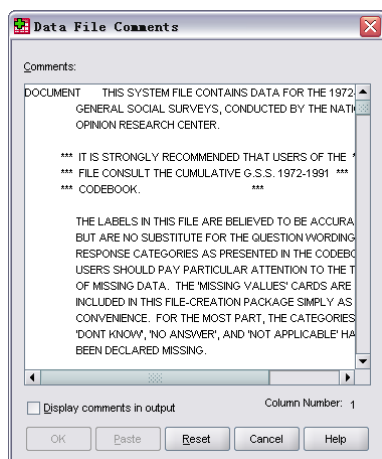


图 2-24 【Data File Comments】对话框

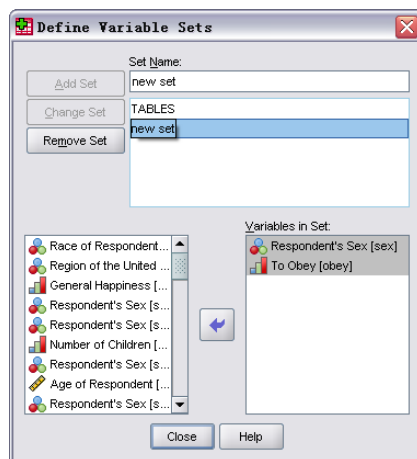


图 2-25 定义变量集对话框

当变量较多且不希望使用所有变量的时候，就可以把要使用的变量单独定义为一个集合。而且此对话框可同时定义多个集合。具体操作方法为：首先将左下方框中要定义为一个集合的变量全部选入右下方框中。然后在“Set Name”框为集合命名，左键单击【Add Set】按钮，则可以看见集合已经添加到上方第二个框图之中。重复此操作就可以定义多个集合。【Change Set】按钮用来为定义好的集合重命名或添加删除变量。【Remove Set】按钮用来取消已定义的集合。

注意 此时一个变量可以同时定义到多个集合之中!

6. 【Use Variable Sets】

也许用户定义了集合之后重新进行统计分析时会觉得定义的集合完全没有发挥作用。这是由于没有执行【Use Variable Sets】命令。执行【Utilities】/【Use Variable Sets】命令，弹出如图 2-26 所示对话框，可以发现刚才新定义的“new set”此时其实还没有发挥作用。需要将“ALLVARIABLES”和“NEWVARIABLES”两项的前面钩去掉，再选中用户自己定义的新变量才算大功告成。

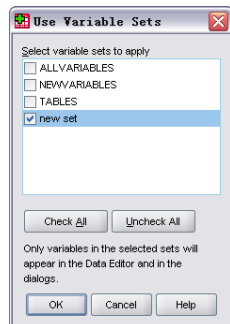


图 2-26 【Use Variable Sets】对话框

7. 【Run Script】

主要针对 SPSS 的二次开发，一般不用。

2.4 本章小结

通过本章的学习，读者需要重点掌握以下知识：

- 在 SPSS 中定义新变量并录入数据。
- 【File】菜单：通过四种方法打开数据文件，数据保存与选择性保存。
- 【Edit】菜单：数据和变量的复制粘贴、插入和查找，其中变量的复制粘贴是一项新增的特殊功能。
- 【Utilities】菜单：定义和使用变量集。

第 3 章 SPSS 数据文件的整理

通过上一章的学习，读者已经知道了如何建立与编辑 SPSS 数据文件。然而，这里并不急于介绍如何利用数据文件进行统计分析，因为在此之前，还有一个很重要的步骤，就是数据文件的整理。这一章就来介绍 SPSS 数据文件的整理。本章的内容包括：

- 数据文件整理概述
- 数据文件的整理——Data 菜单详解
- 变量的变换和计算——Transform 菜单详解

3.1 数据文件整理概述

在建立好数据文件以后，往往还要进行数据文件的加工、整理，经过整理以后的文件才能更好地满足统计分析的要求，这项工作统计学中被称为统计整理。

3.1.1 数据文件的整理在实际工作中的重要性

刚建立的数据文件往往还很粗糙。比如，文件中的数据还是无序的状态，需要对文件的数据按照某个变量或某几个变量来排序，又或者需要根据某几个变量来对文件进行分组，以方便各个组之间进行分析结果的对比等。这些都属于文件整理范围内的事情。SPSS 专门提供了两个菜单来做这些事情，分别是【Data】菜单和【Transform】菜单。其中【Data】菜单主要针对数据文件和观测量的整理，里面包含了 19 个过程；【Transform】菜单则是针对变量的整理，里面包含了 14 个过程。可见 SPSS 17.0 在数据文件整理方面提供的强大功能，也足以说明数据文件整理的重要性。下面我们举一个实例来介绍数据文件的整理。

3.1.2 一个数据文件整理的案例

例 3.1 假设一个公司甲，它有很多的子公司 A、B、C……，分别处在全国的各大城市。每个子公司都有各自的销售记录，保存在各自的数据文件中，也就是说一个子公司对应着一个销售记录信息文件。到了年底，总公司想要总结全年的销售业绩，不仅要各个子公司的业绩加起来统计，还要将各个子公司业绩分别做对比。

各个子公司将自己的销售记录文件交到总公司，总公司首先将这些数据文件统一在一个数据文件中，这时，就可以通过【Data】菜单的【Merge File】过程来实现，具体用法可

参见 3.2.4 节。将这个大的数据文件归并好了以后，又需要将子公司进行分组，对每组的销售业绩进行计算和比较。这个功能可以由【Data】菜单的【Aggregate】过程或【Split File】过程来实现，具体参见 3.2.5 节和 3.2.6 节。

3.2 数据文件的整理——Data 菜单详解

本节将介绍如图 3-1 所示的【Data】菜单。这个菜单正如它的名字所示的那样，里面所有的过程都是针对数据整理的。下面详细介绍它们的功能。

这里说明一下，在以下的介绍中，如果没有特别指明，都是以光盘中的数据文件“Employee Data.sav”为例进行具体的讲解。读者可以自己先打开这个文件，了解里面各个变量的具体含义。

3.2.1 观测量排序——Sort Case 过程

对观测量按照某个或者某几个变量进行排序，是数据文件整理的一项重要内容。SPSS 通过【Data】菜单中的【Sort Cases】过程来实现这一功能。

执行【Data】/【Sort Cases】命令后，弹出如图 3-2 所示的【Sort Cases】对话框，下面就对一一介绍对话框里的重要元素。

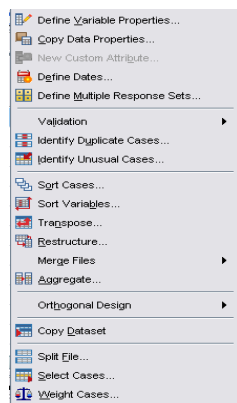


图 3-1 【Data】菜单

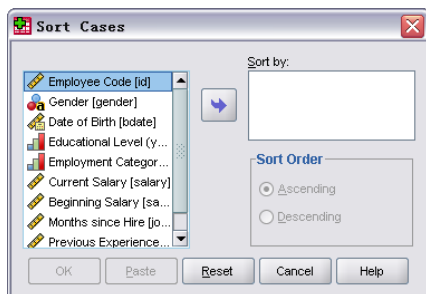


图 3-2 【Sort Cases】对话框

1. 原变量列表框

位于对话框的左侧。里面列出了原文件内的所有变量。

2. Sort by 框

排序变量框。被选入这个列表框内的变量，将作为观测量排序的依据变量，第一个从原变量列表中被选入的变量，作为第一排序变量，数据文件将首先按这个变量进行排序。这时，对于第一排序变量取相同值的变量，系统就按第二个排序变量进行排序，以此类推。

注意 对于字符串变量中的同一个字母，按大写字母优先于小写字母排序。

3. Sort Order单选框

这里的两个选项 Ascending 和 Descending 的意思分别是按升序排列和按降序排列。下面举例说明它们的用法。

例 3.2 要求将数据文件“Employee Data.sav”按照变量 jobcat（工作种类）降序排列，对于相同的值再按照变量 jobtime（参加工作月数）升序排列。具体步骤如下。

执行【Data】/【Sort Cases】命令，弹出【Sort Cases】对话框		
【Sort by】：jobcat	选择 jobcat 作为第一排序变量	
在 Sort Order 单选框：Descending	按照 jobcat 变量降序排列	
【Sort by】：jobtime	选择 jobtime 作为第二排序变量	
在 Sort Order 单选框：Ascending	按照 jobtime 变量升序排列	
单击【OK】按钮	新数据文件覆盖原文件	

3.2.2 数据文件转置——Transpose过程

数据文件转置，也就是将数据文件中的行、列进行互换。利用【Data】菜单的【Transpose】过程，可将观测量转化为变量，将变量转化为观测量。

执行【Data】/【Transpose】命令，弹出如图 3-3 所示【Transpose】对话框，首先介绍其中的重要元素。

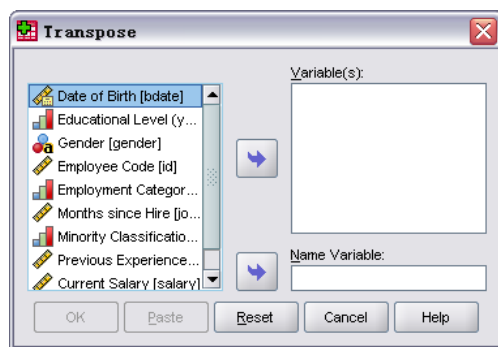


图 3-3 【Transpose】对话框

1. 原变量列表框

位于对话框左侧。其中列出了所有的原文件变量。

2. Variable框

变量框。将需要进行转置的变量移入其中。

3. Name Variable框

变量命名框。在原变量列表框中选择一个变量移入这个框内，这个变量就作为转置后的新变量名。一般选择有命名作用的变量，比如姓名、编号等等。如果没有变量移入这个框，系统将自动生成新变量 Var001、Var002、……。

注意 字符型变量不能进行转置，如果强行转置，就会输出缺失值。

下面举例说明它的用法。

例 3.3 要求将 salary（工资）、educ（受教育年数）、jobcat（工作种类）进行转置，并且以 id（序号）为新变量名。操作步骤如下。

执行【Data】/【Transpose】命令后，弹出【Transpose】对话框

【Variable】: salary、educ 和 jobcat

准备将变量 salary、educ 和 jobcat 进行转置

【Name Variable】: id

将 id 中观测值作为新变量名

单击【OK】按钮

出现一个提示信息，提示用户其他未被选择转置的变量其数据将会丢失

执行以上操作之后，将新生成一个如图 3-4 所示的数据文件。

	CASE_LBL	K_1	K_2	K_3	K_4	K_5
1	salary	57000.00	40200.00	21450.00	21900.00	45000.00
2	educ	15.00	16.00	12.00	8.00	15.00
3	jobcat	3.00	1.00	1.00	1.00	1.00

图 3-4 转置后的新数据文件

3.2.3 数据格式重排——Restructure过程

顾名思义，数据格式重排即根据用户需要，重新改变数据的排列格式。比如在多次重复的实验中，将同一个体多次实验的数据结果转变为多行观测量分别显示，或者将同一个体的多次多行观测量放到一行观测量中显示。

执行【Data】/【Restructure】命令，弹出如图 3-5 所示的对话框。该对话框有 3 个单选项，分别对应 3 种形式的数据结构重排。

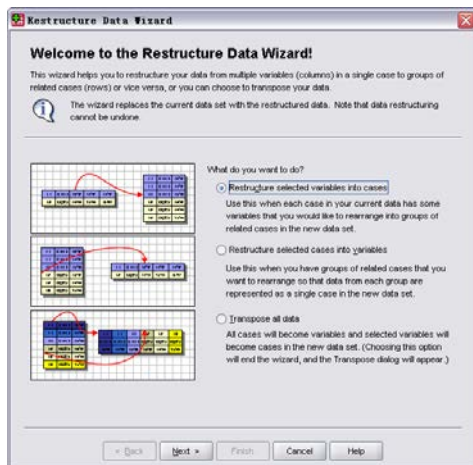


图 3-5 【Restructure Data Wizard】对话框

- 第一个选项的意思是将一行观测量的多个变量转换为相关的多行观测量，即将宽型数据转变为长型数据；
- 第二个选项的意思是将相关的多行观测量转换为一行观测量的多个变量，即将长型数据转变为宽型数据；
- 第三个选项的意思是将所有变量行列转置，即刚刚介绍的 Transpose 功能。实际上，如果选择这一项的话，再单击【Next】按钮，就会出现【Transpose】对话框。

读者可以看到，在这 3 个选项的左侧都有这个选项所对应的图解，可以清晰地表达这个选项的意义。

因为这个功能是通过多组对话框不断询问格式重排的要求来实现这个功能的，这里就不再一一介绍每个对话框了。举一个实际的例子来帮助读者了解它的用法。

例 3.4 设某校对英语专业的同学分别进行了听、说、读、写四次考试，每次考试成绩分别单独记录在如图 3-6 所示的数据文件“exam.sav”中。利用【Restructure】命令，将同一考生的四门考试成绩用一条记录显示。

STEP 01 在【Restructure Data Wizard】对话框中选择第二个选项，然后单击【Next】

	name	Scores	Subject	var
1	Andy	18.00	writing	
2	Andy	14.00	Listening	
3	Andy	12.00	reading	
4	Andy	6.00	speaking	
5	Tracy	19.00	writing	
6	Tracy	12.00	Listening	
7	Tracy	8.00	reading	
8	Tracy	4.00	speaking	
9	Nick	14.00	writing	
10	Nick	10.00	Listening	
11	Nick	6.00	reading	
12	Nick	2.00	speaking	

图 3-6 转换前的长型数据结构

按钮，出现如图 3-7 所示的第 2 步对话框。该对话框用来设置重复测量的个体的 ID 变量以及区分重复测量次数的 Index 变量。在本例中，选入变量“Name”作为 ID 变量，选入变量“Subject”作为 Index 变量。

STEP 02 当上一步设置完成之后，单击【Next】

按钮，弹出如图 3-8 所示的第 3 步对话框。该对话框用来选择是否对 ID 变量和 Index 变量进行排序。这是因为在数据结构重排之前，ID 变量和 Index 变量必须是排好序的。如果对数据不太了解的话，通常都选择“Yes”。

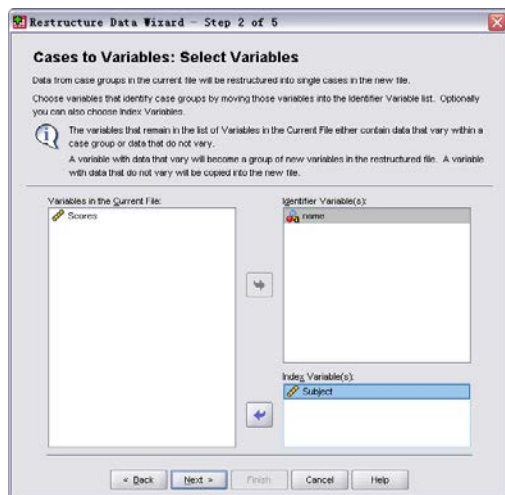


图 3-7 Restructure Data Wizard 的第 2 步

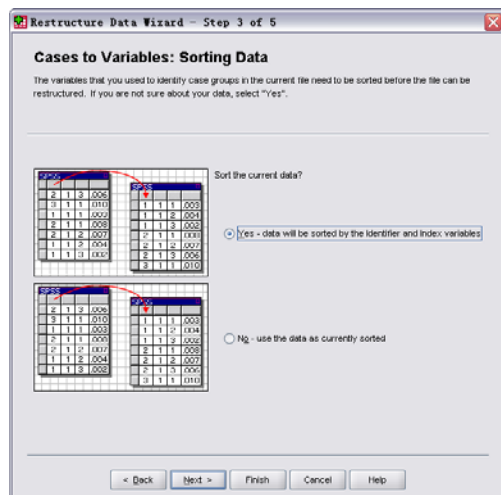


图 3-8 Restructure Data Wizard 的第 3 步

STEP 03 设置完成之后,单击【Finish】按钮,生成如图 3-9 所示的新数据文件。此时原来的长型数据文件已经重排为宽型数据文件。

	name	Scores.1.00	Scores.2.00	Scores.3.00	Scores.4.00
1	Andy	18.00	14.00	12.00	6.00
2	Echo	16.00	12.00	10.00	4.00
3	Nick	14.00	10.00	6.00	2.00
4	Tracy	19.00	12.00	8.00	4.00

图 3-9 格式重排以后的新数据文件

3.2.4 数据文件合并——Merge File子菜单

在实际问题中,经常会需要将不同的数据文件按照行或列合并为一个数据文件,在 SPSS 中称之为数据文件的合并。这个功能是通过【Data】菜单的【Merge File】子菜单来实现的。【Merge File】又包含【Add Cases】和【Add Variables】两个子过程,下面分别举例介绍。

1. 观测量合并——Add Cases过程

例 3.5 已知某公司现将若干名实习生转为正式员工,实习生的资料保存在如图 3-10 所示的一个名为“实习生.sav”的数据文件中,利用【Add Cases】过程将“实习生.sav”数据文件和原始的员工信息“Employee Data.sav”文件合并为一个新数据文件。

	num	name	sex	bdate	educ	jobcat	salary	jobdata
1	475	陈婷婷	female	08/16/1984	16	2	\$1,500	02/17/2006
2	476	兰宇	male	04/25/1982	15	1	\$1,000	11/11/2005
3	477	蒋捷	male	05/28/1981	16	2	\$1,500	02/20/2006
4	478	高燕	female	03/21/1983	15	1	\$1,000	11/12/2005
5	479	刘玲	female	06/15/1983	15	1	\$1,000	11/15/2205
6	480	赵斌	male	04/05/1983	17	2	\$1,500	03/01/2006

图 3-10 “实习生.sav”文件

操作步骤如下:

STEP 01 打开数据文件“Employee Data.sav”作为当前数据文件,执行【Data】/【Merge File】/【Add Cases】命令,弹出如图 3-11 所示的【Add Cases】对话框。如果此时外部文件“实习生.sav”已被打开,就会显示在“An open Dataset”框中,否则可以单击“An external SPSS Statistics data file”框旁边的【Browse】按钮,找到这个数据文件所在的位置。选好以后,单击【Continue】按钮。

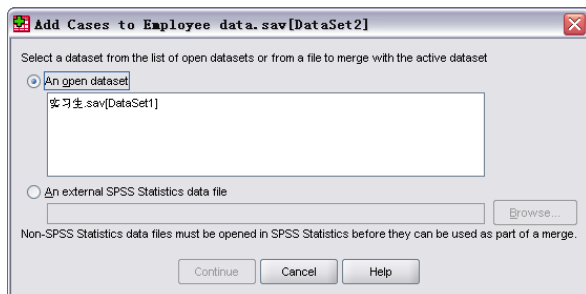


图 3-11 【Add Cases】对话框

STEP 02 弹出如图 3-12 所示的对话框，先来介绍其中的元素。

• Variables in New Active Dataset 框

新数据文件变量框。在其中列出了合并之后生成的新文件将有的变量名。系统默认将当前文件和外部数据文件中名称相同的变量列入其中。如果两个变量的名字相同，但意义不同，可以通过箭头按钮将其移到“Unpaired Variables”框。

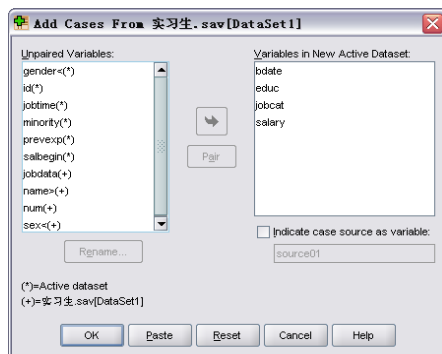


图 3-12 【Add Cases from】对话框

• Unpaired Variables 框

不配对变量框。其中列出了两个文件中所有不同名的变量名。标记“[*]”的变量属于当前文件，标记“[+]”的变量属于外部文件，若变量名后有符号“<”，表示此变量为字符型变量。在此栏下面有一个【Rename】按钮，单击它可以对选中的变量重命名。在实际问题中还常常遇到以下情况，两个变量意义相同，数据类型相同，但名称不同，这时可以单击箭头按钮下方的【Pair】按钮将两者配对，配对后的变量自动移入 Variables in New Active Dataset 框。

注意 若被配对的两个变量类型不同，则这个变量将不会出现在新文件中。

• indicate case source as variable 选项

如果选择这个选项，则系统将在新文件中创建一个新变量，用来标记观测量是来自哪个文件的，且系统默认的变量名为 source01。

这里要求将变量 id 和 num 配对，变量 gender 和 sex 配对，将外部文件的变量 salary 改名为 salbegin，然后与当前文件变量 salbegin 配对，并将外部文件的 jobdata 变量放入新数据文件中，保留当前文件的所有变量。具体步骤如下：

【pair】：id 和 num

【pair】：gender 和 sex

Unpaired Variables 栏：salary

【Rename】：salary[+] salbegin[+]

【pair】：salbegin[+]和 salbegin[*]

Variables in New Active Dataset 栏：所有带[*]变量 将所有原文件变量移入 Variables in New Active Dataset 框

配对变量 id 和 num

配对变量 gender 和 sex

将变量 salary 移入 Unpaired Variables 框

将外部文件变量 salary 改名为 salbegin

配对两个文件的 salbegin 变量

将所有原文件变量移入 Variables in New Active Dataset 框

Variables in New Active Dataset 栏：jobcat[+] 将外部文件变量 jobcat 也移入
单击【OK】按钮 生成如图 3-13 所示新数据文件

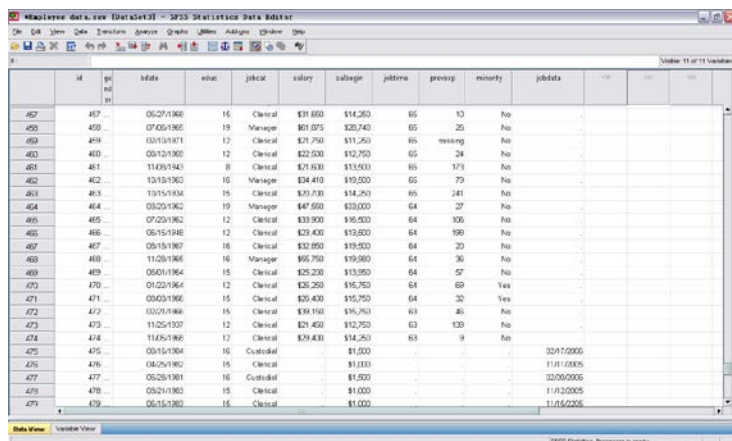


图 3-13 增加观测量以后的新文件

由于原始文件“Employee Data.sav”中没有变量 jobdata，因此在新数据文件中，属于“Employee Data.sav”的观测量在此变量下的值都是缺失值。

2. 变量合并——Add Variables过程

例 3.6 已经有一如图 3-14 所示的外部数据文件“manager.sav”，该数据文件表示员工是否为经理。利用【Add Variables】过程，合并外部数据文件“manager.sav”到当前数据文件“Employee Data.sav”中。

操作步骤如下：

STEP 01 打开当前数据文件“Employee Data.sav”，执行【Data】/【Merge File】/【Add Variables】命令，弹出如图 3-15 所示【Add Variables】对话框，以类似于图 3-11 的方法，选好数据文件以后，单击【Continue】按钮。

	num	sex	vacation
1	1	m	0
2	18	m	0
3	27	m	1
4	29	m	0
5	32	m	0
6	34	m	0
7	35	m	0
8	50	m	0
9	53	m	0
10	62	m	1
11	63	m	0
12	64	m	0
13	66	m	0
14	67	m	0
15	68	m	0

图 3-14 “manager.sav”数据文件

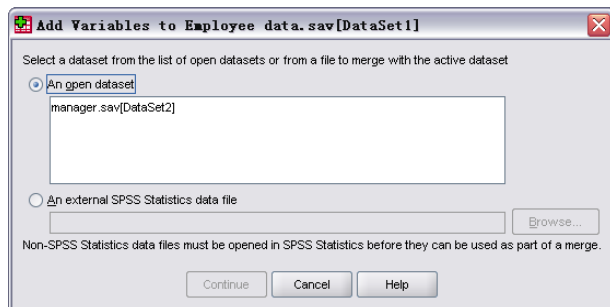


图 3-15 【Add Cases】对话框

STEP 02 弹出如图 3-16 所示的对话框，先来介绍其中的元素。

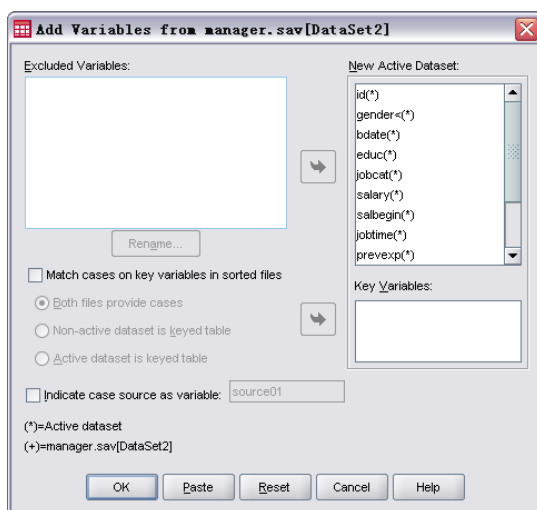


图 3-16 【Add Variables from】对话框

- New Active Dataset 框

新数据文件框。其中列出了合并之后的新数据文件将有的所有变量。系统默认为两个文件中的所有不同名的变量。

- Excluded Variables 框

被排除变量框。其中列出了所有被 New Active Dataset 框排除的变量。若想对其中变量改名，选中该变量之后，单击下面的【Rename】按钮。

- Match cases on key variables in sorted files 选项

如果不选择该项，则 SPSS 默认对两个数据文件中记录号相等的观测量进行合并。选择该项，SPSS 则将关键变量相等的观测量进行合并，其中包含 3 个单项：

第一个选项的含义是合并两个文件关键变量值相等的观测量到新文件中，对于两文件中没有与之关键变量相等配对的观测量，分别各自作为一条单独的记录纳入新数据文件之中；

第二个选项的含义是合并两个文件关键变量值相等的观测量到新文件中，对于两文件中没有与之关键变量相等配对的观测量，只保留当前数据文件中所有的观测量；

第三项的含义是合并两个文件关键变量值相等的观测量到新文件中，对于两文件中没有与之关键变量相等配对的观测量，只保留外部数据文件中所有的观测量。

- Key Variable 框

其中放置将两个文件联系起来的关键变量。如果关键变量在两个文件中名称不同，首先在 Excluded Variables 框对其中一个改名，使其名称相同后选入 Key Variable 框。

注意 关键变量必须是两个文件都共有的，且类型一致，分类排序顺序一致。同时，关键变量最好选择有唯一标识性的变量，如姓名、编号等等。

- indicate case source as variable 选项

如果选择这个选项，则系统将在新文件中创建一个新变量，用来标记观测量是来自哪个文件的，系统默认的变量名为 source01。

这里要求合并后的文件只保留原文件和外部文件中关键变量值相等的观测量，保留“manager.sav”文件（这里作为外部数据文件）中所有的观测量。具体步骤如下：

选择“Match cases on key variables in sorted files”

单选“active dataset is keyed table”

Excluded Variables 栏：gender、id 和 num

【Rename】：id[*]->num[*]

Keyed Variable 栏：num

单击【OK】按钮

两个文件中观测量数目不同

保留“manager.sav”中观测量

将这三个变量移入 Excluded Variables 栏

将变量 id 改名为 num

任选一个变量 num 移入 Keyed Variable 栏

新的数据文件如图 3-17 所示

	num	bdate	educ	jobcat	salary	salbegin	jobtime	prevexp	minority	sex	vacation
1	1	02/03/1952	15	Manager	\$57,000	\$27,000	98	144	No male	no	
2	18	03/20/1956	16	Manager	\$103,750	\$27,510	97	70	No male	no	
3	27	03/19/1954	19	Manager	\$60,375	\$27,480	96	96	No male	yes	
4	29	01/28/1944	19	Manager	\$135,000	\$79,980	96	199	No male	no	
5	32	01/28/1954	19	Manager	\$110,625	\$45,000	96	120	No male	no	
6	34	02/02/1949	19	Manager	\$92,000	\$39,990	96	175	No male	no	
7	35	08/22/1961	17	Manager	\$81,250	\$30,000	96	18	No male	no	
8	50	03/09/1960	16	Manager	\$60,000	\$23,730	94	59	No male	no	
9	53	04/21/1954	18	Manager	\$73,750	\$26,250	94	56	No male	no	
10	62	07/18/1962	16	Manager	\$49,000	\$21,750	93	22	No male	yes	
11	63	06/20/1961	17	Manager	\$55,000	\$26,250	93	32	No male	no	
12	64	08/28/1963	16	Manager	\$53,125	\$21,000	93	48	No male	no	
13	66	02/16/1962	19	Manager	\$78,125	\$30,000	93	7	No male	no	
14	67	05/28/1964	16	Manager	\$45,000	\$21,240	93	35	No male	no	
15	68	05/05/1963	16	Manager	\$45,250	\$21,480	93	36	No male	no	
16	69	05/23/1960	16	Manager	\$56,550	\$25,000	93	34	No male	no	
17	71	08/26/1948	17	Manager	\$82,500	\$34,980	93	207	No male	no	
18	88	02/10/1962	19	Manager	\$72,500	\$26,740	92	10	No male	no	
19	89	06/24/1961	19	Manager	\$68,750	\$27,480	92	8	No male	no	
20	100	10/25/1963	18	Manager	\$78,250	\$27,480	91	47	No male	no	
21	101	03/14/1960	16	Manager	\$60,625	\$22,500	91	44	No male	no	
22	103	03/17/1959	19	Manager	\$97,000	\$35,010	91	68	No male	no	
23	106	08/04/1962	19	Manager	\$91,250	\$29,490	91	23	No male	no	
24	113	10/06/1959	16	Manager	\$54,875	\$27,480	90	68	No male	no	
25	120	11/12/1964	16	Manager	\$37,800	\$15,750	90	7	No female	no	

图 3-17 增加变量以后的新文件

如图 3-17 所示，新生成的数据文件中就多了一列变量“Vacation”。

3.2.5 数据分类汇总——Aggregate过程

数据分类汇总功能能够将观测量按照某几个变量进行分组，并根据分组对某些变量求其分类汇总值。

执行【Data】/【Aggregate】命令，弹出如图 3-18 所示的【Aggregate Data】对话框，先来介绍对话框的主要元素。

1. 原变量列表框

相信读者应该对它已经很熟悉了，这里不再赘述。

2. Break Variable框

分组变量框。其中放置分组变量，即将观测量进行分组的依据变量。

3. Summaries of Variable框

汇总变量框，其中放置待汇总的变量。

4. Function按钮

定义各个汇总变量的汇总函数。单击【Function】按钮，弹出如图 3-19 所示的【Aggregate Function】（汇总函数）对话框。该对话框的每个选项代表了一个汇总函数，它们的意义如表 3-1 所示。

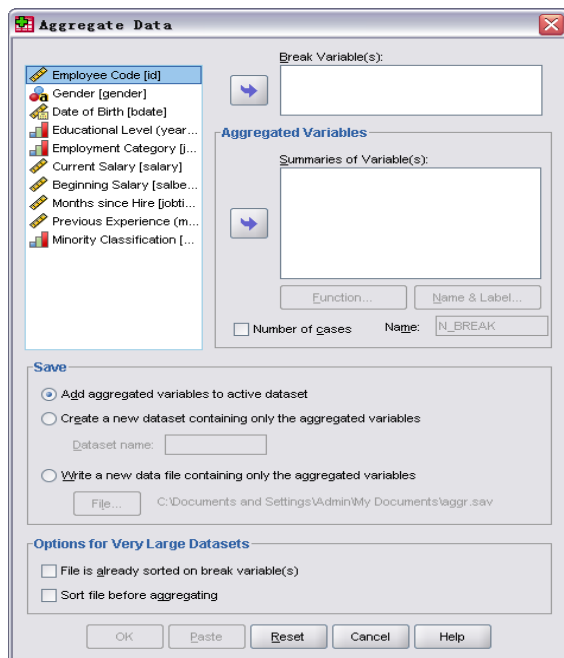


图 3-18 Aggregate Data 主对话框

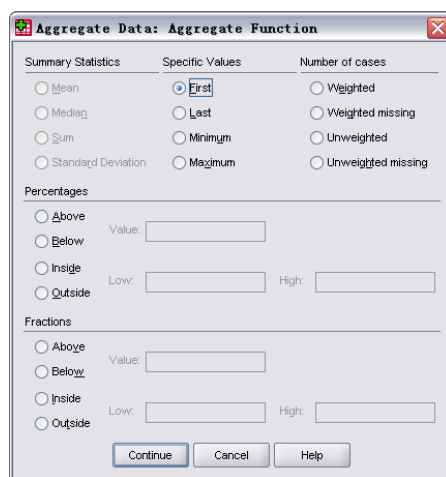


图 3-19 汇总函数对话框

注意 汇总函数只针对数值型变量而言。

表 3-1 汇总函数

栏 名	选 项 名	选项意义
Summary (概括函数)	Mean	算术平均数，这是系统默认值
	Median	中位数
	Sum	观测值之和
	Standard deviation	观测值的标准差
Specific Value (特殊值)	First	第一个观测值
	Last	最后一个观测值
	Minimum	最小观测值
	Maximum	最大观测值

续表

栏 名	选 项 名	选项意义
Number of cases (观测量总数)	Weighted	对加权的文件, 汇总计算原变量的有效值中各分组变量的观测值数
	Weighted missing	汇总计算加权的数据文件中原变量的缺失值数
	Unweighted	对未经加权的文件, 汇总计算原变量的有效观测值中各分组变量的观测值数
	Unweighted missing	汇总计算未加权的数据文件中原变量的缺失值数
Percentage (百分比)	Above	观测值大于指定值的观测值数占全组观测值总数的百分数, 指定值在 Value 栏内定义
	Below	观测值小于指定值的观测值数占全组观测值总数的百分数, 指定值在 Value 栏内定义
	Inside	观测值介于两个指定值之间的观测量数占全组观测值总数的百分数, 指定值在 Low 栏和 High 栏内定义
	Outside	观测值介于两个指定值之外的观测量数占全组观测值总数的百分数, 指定值在 Low 栏和 High 栏内定义
Fractions (小数)	Above	观测值大于指定值的观测值数所占全组观测值总数的比率, 指定值在 Value 栏内定义
	Below	观测值小于指定值的观测值数所占全组观测值总数的比率, 指定值在 Value 栏内定义
	Inside	观测值介于两个指定值之间的观测量数所占全组观测值总数的比率, 指定值在 Low 栏和 High 栏内定义
	Outside	观测值介于两个指定值之外的观测量数所占全组观测值总数的比率, 指定值在 Low 栏和 High 栏内定义

5. Name & Label按钮

单击【Name & Label】按钮, 也会出现一个对话框。该对话框主要用来定义新生成汇总变量的变量名和变量标签。

6. Number of Cases复选框

勾选该选项, 则会在新文件中创建一个变量用来计算分组后每一组观测量的数目, 系统默认该变量的变量名为 N_BREAK。

7. Save单选框

定义新生成汇总变量的保存方法, 共包含 3 个选项, 其含义如下:

- 第一个选项的意思是将新生成的汇总变量加入到原数据文件中;
- 第二项是指建立一个新的数据集, 里面只包含新生成的汇总变量, 并在 Dataset Name 栏内命名这个数据集;
- 第三项是指建立一个新的数据文件, 里面只包含新生成的汇总变量, 单击【File】按钮选择新文件的保存位置和文件名。

8. Options for Very Large Datasets单选框

处理大型数据集文件, 在分类汇总之前需将数据按照分组变量进行排序, 其中:

- 第一个选项意思是文件已经按照分组变量排好序了;

- 第二项是指在汇总之前先将文件排序。

对于大型数据集，通常选择第二项来提高计算机效率。

现在举一个例子来说明【Aggregate】过程的具体操作。

例 3.7 要求将数据文件“Employee Data.sav”按照变量 gender（性别）进行分组，对每一组的变量 salary 计算其算术平均数，对变量 salbegin 计算其最大观测值，对变量 minority 计算其值为 1 的观测量数目占全组观测量总数的百分数，并建立一个新数据文件放置分类汇总以后的结果。操作步骤如下：

【Break Variable】：gender	在【Aggregate Data】主对话框中将变量 gender 作为分组变量
【Summaries of Variable】：salary	选择 salary 作为汇总变量，系统默认计算其均值
【Summaries of Variable】：salbegin	选择 salbegin 汇总变量
单击【Function】按钮	弹出汇总函数对话框
选择 Specific Value 栏内的选项“Maximum”	计算最大值
单击【Continue】按钮	回到【Aggregate Data】主对话框
【Summaries of Variable】：minority	选择 minority 作为汇总变量
单击【Function】按钮	弹出汇总函数对话框
选择 Percentage 栏内的选项“Inside”	
然后在 Low 栏内、High 栏内均输入 1	计算取值为 1 的观测量所占百分比
单击【Continue】按钮	回到【Aggregate Data】主对话框
在 Save 单选框内选择第三项	这时系统默认文件名为“aggr.sav”
单击【OK】按钮	定义完成

分组汇总后的结果出现在新文件中，如图 3-20 所示。

	gender	salary_mean	salbegin_max	minority_pin	
1	Female	26031.92	\$30,000	18.5	
2	Male	41441.78	\$79,980	24.8	

图 3-20 分组汇总后的新数据文件

通过新生成的数据文件可知，女性雇员的薪水平均值以及起薪最大值都远低于男性雇员。男性雇员中少数民族的占比略大于女性雇员。

3.2.6 数据文件的拆分——Split File过程

数据文件的拆分是指将数据文件按某个或某几个变量取值拆分为多个独立的数据文件。在数据文件拆分完成以后，数据文件的显示虽然和拆分前差别不大，但是已经可以将其看成是多个独立的数据文件了。数据文件拆分是【Data】菜单中最常用和最实用的一项功能。

执行【Data】/【Split File】命令，弹出如图 3-21 所示的【Split File】对话框。现在介绍它的重要组成元素。

1. Groups Based on 框

被移入其中的变量就是拆分变量，即文件拆分的依据变量。

2. 文件拆分单选框

位于 Groups Based on 框的上面，有如下 3 个选项。

- **Analyze all cases, do not create groups**，这是默认选项，表示不对数据文件进行拆分，可利用该选项取消数据文件的拆分；
- **Compare groups**，若选择这一项，在统计分析的时候，各个拆分文件的输出结果将放在同一个表格中相互比较；
- **Organize output by groups**，若选择这一项，在统计分析的时候，各个拆分文件的输出结果将分别用一个独立的表格单独输出。

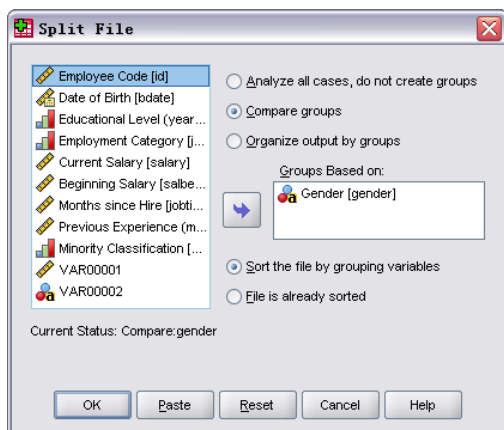


图 3-21 【Split File】对话框

3. 排序单选框

在拆分文件之前必须对原始的观测量按照拆分变量进行排序，排序单选框有两个选项：

- **Sort the file by grouping variables**，将数据文件按照拆分变量进行排序；
- **File is already sorted**，数据文件已经排好序了。

此时通常选择第一项。

下面举例说明【Split File】过程的使用。

例 3.8 要求以 gender（性别）对数据文件“Employee Data.sav”进行拆分，并要求在以后的统计分析中可以将各拆分文件的统计分析结果放在同一表格中显示。操作步骤如下：

选择文件拆分单选框中的“Compare groups”按钮 选择将拆分文件的统计输出结果放在同一表格相互比较

【Groups Bases on】: gender

将变量 gender 作为拆分变量

单击【OK】按钮

数据编辑窗口变为如图 3-22 所示

如图 3-22 所示，拆分后的数据文件乍一看同拆分前相比似乎仅仅是按照拆分变量进行了排序而没有其他区别。细心的读者此时可能已经发现了，拆分后的数据文件右下方多了一行说明“Split by gender”，即此时数据文件已经是拆分了的。这时再进行统计分析的话，输出的统计分析结果将按照性别分别显示。

	id	gender	birthdate	refcat	jobcat	salary	subbegin	jobtime	greenup	tenure
1	3	Female	07/06/1929	12	Clerical	\$21,450	\$12,000	36	381	No
2	4	Female	04/01/1947	8	Clerical	\$21,900	\$11,200	36	140	No
3	5	Female	05/05/1955	12	Clerical	\$21,900	\$9,750	36	missing	No
4	9	Female	01/23/1946	15	Clerical	\$27,900	\$12,750	36	115	No
5	13	Female	02/15/1946	17	Clerical	\$24,000	\$11,600	36	264	No
6	11	Female	02/07/1950	16	Clerical	\$30,300	\$15,500	36	143	No
7	14	Female	02/26/1949	15	Clerical	\$35,100	\$15,800	36	137	Yes
8	23	Female	01/02/1940	17	Clerical	\$26,760	\$11,500	37	48	No
9	21	Female	02/15/1953	16	Clerical	\$38,850	\$15,000	37	17	No
10	23	Female	03/15/1965	15	Clerical	\$24,000	\$11,100	37	75	Yes
11	24	Female	03/27/1939	17	Clerical	\$15,960	\$9,000	37	134	Yes
12	25	Female	07/01/1942	15	Clerical	\$21,150	\$9,000	37	171	Yes
13	36	Female	05/07/1963	8	Clerical	\$31,350	\$11,200	36	52	No
14	43	Female	05/25/1933	15	Clerical	\$19,200	\$9,000	36	23	Yes
15	41	Female	02/15/1951	12	Clerical	\$25,550	\$11,550	36	52	Yes
16	46	Female	11/16/1940	15	Clerical	\$22,350	\$12,750	36	195	No
17	47	Female	04/26/1938	12	Clerical	\$30,000	\$15,600	36	239	No
18	50	Female	11/14/1954	15	Clerical	\$25,400	\$13,500	34	3	No
19	72	Female	01/08/1964	16	Clerical	\$54,000	\$15,000	33	11	No
20	73	Female	02/05/1958	12	Clerical	\$25,400	\$13,500	33	missing	No
21	74	Female	04/26/1933	15	Clerical	\$33,000	\$13,500	33	132	No
22	75	Female	08/12/1965	15	Clerical	\$24,150	\$11,550	33	missing	No
23	76	Female	05/03/1967	16	Clerical	\$29,250	\$11,550	33	11	No
24	77	Female	05/05/1950	12	Clerical	\$27,600	\$11,400	33	6	No
25	73	Female	08/22/1966	12	Clerical	\$22,950	\$13,500	33	19	No

图 3-22 拆分后的数据文件

注意 数据文件一旦拆分之后，如果没有使用【Split File】过程取消拆分的话，那么在以后所有的统计分析过程中，拆分都一直存在。

3.2.7 选择观测量——Select Cases过程

选择观测量过程是指从数据文件中选取符合要求的观测量作为样本参与统计分析。在实际问题中，这也是一个经常用到的功能。

执行【Data】/【Select Cases】命令，弹出如图 3-23 所示的【Select Cases】对话框，下面就来介绍它的各个元素。

1. Select 单选框

关于观测量选择的单选框。下面分别讲解每个选项的含义。

• All cases 选项

系统默认值，意思是选择全部观测量。

• If condition is satisfied 选项

选择满足条件的观测量。

单击【If】按钮，弹出如图 3-24 所示的【If】子对话框。对话框右上方的空白栏是表达式栏；下面类似计算器界面的框里是 SPSS 的所有运算符，关于这些运算符的含义见本章 3.3.1 节的表 3-3；旁边是 Function group 列表框和 Functions and Special Variables 框（函数组列表框和函数与特殊变量框）。Function group 列表框里列举了 SPSS 的所有函数组，单击任意一个组名，这一组中所有的函数和特殊变量将出现在 Functions and Special Variables 框内，再单击任意一个函数名，这个函数的信息就出现在计算器板下面的空白框

中，供用户查询这个函数的意义和用法。这样的设计与老版本 SPSS 中只有一个 Function 框列举出所有函数的方法相比，更加方便。

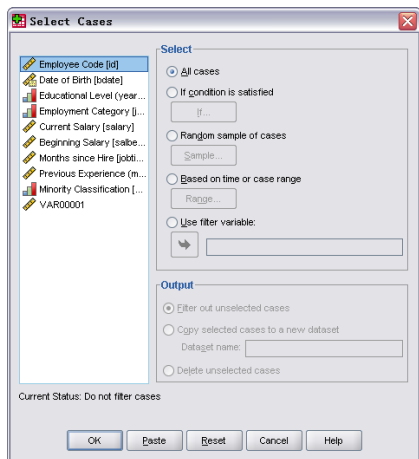


图 3-23 【Select Cases】对话框

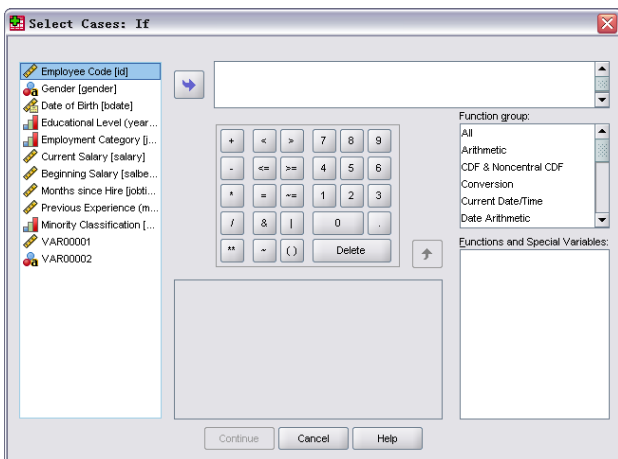


图 3-24 If 子对话框

比如，我们要在“Employee Data.sav”文件中选择现在的工资比开始工作时的工资涨了 10000 以上的员工，这时就在表达式栏内输入表达式“(salary - salbegin) >= 10000”即可。这时系统自动在文件中增加一个名为“filter_\$”的变量，用来标记该观测量是否被选中，选中的观测量对应值为 1，否则为 0。

- Random sample of cases 选项

随机抽取观测量样本。单击【Sample】按钮，弹出如图 3-25 所示的【Random Sample】子对话框，其中：

第一个选项的意思是随机选取占全部观测量的百分比的观测量，在%号前输入这个百分比；

第二项意思是从前面多少个观测量中随机选择多少个观测量数。

- Bases on time or case range 选项

按观测量记录号选择。单击【Range】按钮，弹出如图 3-26 所示【Range】子对话框，若在 First Case 栏里填入 20，在 Last Case 栏里填入 50，则系统将选择观测量中记录号从 20 到 50 的观测量。

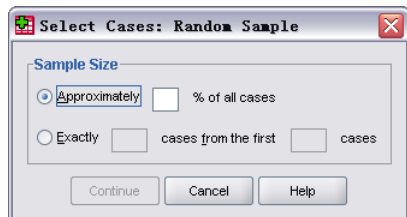


图 3-25 【Random Sample】子对话框



图 3-26 【Range】子对话框

- Use filter variable 选项

使用过滤器变量选择观测量。从变量列表框中选中的一个变量移入 Use filter variable 栏，系统将依据这个变量，将该变量值为 0 或缺失的观测量过滤掉，而选择该变量取值不为 0 的观测量。

2. Output 单选框

关于输出结果选择的单选框。有 3 个单选项：

- 第一项意思是过滤未被选中的观测量，即未被选中的观测量仍然保留在数据文件中，只是其对应的观测量序号上划上了斜线“\”作为标记，这些观测量将在以后进行的统计分析中暂时被关闭；
- 第二项意思是将选择好的观测量复制到一个新的数据集里，并在下面的 Dataset name 栏里命名这个数据集；
- 第三项意思是从原文件中删除未被选中的观测量。

3.2.8 观测量加权——Weight Cases 过程

在前面介绍汇总函数时，出现了加权文件和未加权文件的概念，下面来讲解它的意思。权重是指同一个观测量值在所有的观测量里出现的次数或者频率。SPSS 的观测量加权功能是在数据文件中选择一个变量，这个变量取值就代表相应观测量出现的次数，这个变量叫做权变量，经过加权的数据文件叫做加权文件。以数据文件“smoking.sav”为例，其中的变量“count”是指拥有相等的职位，又有相等的吸烟程度的人的个数，因此，可以利用【Weight Cases】过程将这个变量定义为权变量。

执行【Data】/【Weight Cases】命令，弹出如图 3-27 所示的【Weight Cases】对话框。该对话框上面有一个单选框，其中：

- 第一个选项意思是不对观测量加权，可用该选项来取消对变量的加权；
- 第二项意思是对观测量加权，在 Frequency Variable 栏内选入一个变量作为权变量。

选好以后单击【OK】按钮即可。同【Split File】过程类似加了权的数据文件在数据编辑窗口中与未加权时相比似乎完全一样，只是在数据编辑窗口的右下方多了一行说明“Weight on”。

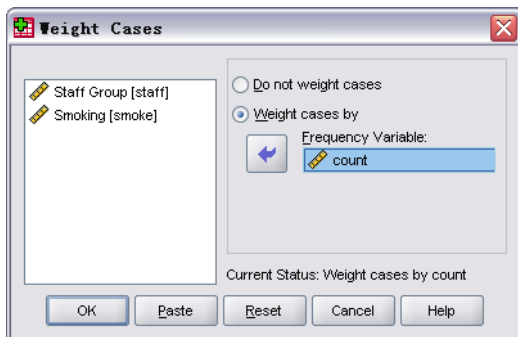



图 3-27 【Weight Cases】对话框

 **注意** 如果没有使用【Weight Cases】过程取消加权的话，那么在以后所有的统计分析过程中，观测量加权都一直存在。

3.2.9 Data菜单其他过程简介

上面介绍了 Data 菜单中几个常用的数据文件整理的功能。剩下的一些特殊功能简要介绍如表 3-2 所示。

表 3-2 Data 菜单其他过程简介

名 称	作 用
【Define Variable Properties】	定义选定变量的变量性质
【Copy Data Properties】	以一个外部 SPSS 文件为模板（也可以是当前文件），来为活动数据集定义其中选定变量的变量值标签或者数据集的性质
【Define Datas】	自动生成时间变量，主要用于时间序列模型
【Define Multiple Response Sets】	复选变量集，用于定义 Multiple Response Variables（复选变量），它是由 Custom Tables and the Chart Builder 所支持的一种特殊的“变量”，详见 5.2.1 节
【Validation】	主要用于定义单个变量或者交叉变量的有效性，以及检验变量值的有效性
【Identify Duplicate Cases】	识别重复的观测量
【Identify Unusual Cases】	识别异常的观测量
【Sort Variables】	对变量排序
【Orthogonal Design】	用于自动生成正交设计表格，用于在统计分析中做正交设计，比如一次回归分析中的正交设计
【Copy Dataset】	为当前工作数据文件复制一个新的数据集作为模板，用于定义文件或者变量的性质

3.3 变量的变换和计算——Transform菜单详解

上一节介绍了针对数据文件和观测量的整理——Data 菜单，这一节介绍针对变量的整理——Transform 菜单，如图 3-28 所示，下面介绍它的每个过程的功能。

3.3.1 变量计算——Compute Variable过程

SPSS 提供了强大的变量计算功能。刚刚建立的数据文件中的数据还都是原始数据，而统计分析需要利用这些实际测量的原始数据，找出变量之间的关系，进而揭示现象的内在数量规律。利用原始数据计算一些特定的数值，组成一个新的变量，就是【Compute Variable】过程所提供的功能。

执行【Transform】/【Compute Variable】命令，弹出如图 3-29 所示的【Compute Variable】对话框，首先介绍其中的主要元素。



图 3-28 【Transform】菜单

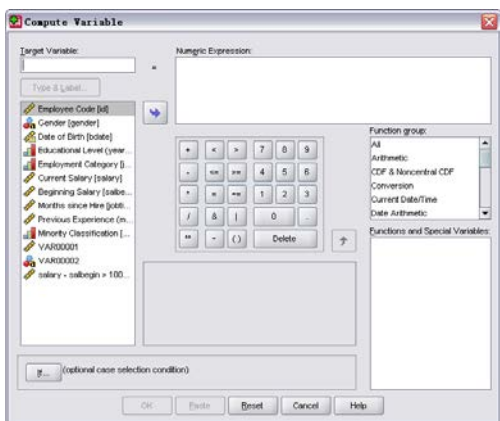


图 3-29 【Compute Variable】对话框

1. Target Variable框

目标变量框，在这个框内定义将要产生的目标变量。在空白栏内输入目标变量的名称，可以是一个新变量名，也可以是已经定义的变量名（计算完成后原观测值将被覆盖）。单击【Type & Label】栏，出现一个子对话框，在这个子对话框内可以定义目标变量的 Label（标签）和 Type（数据类型）。

2. Numeric Expression栏

数值表达式栏，设置新变量的计算表达式。

3. 计算器板

对话框中类似计算器界面的框，其中的按钮就是 SPSS 的所有运算符，各运算符的意义如表 3-3 所示。

表 3-3 SPSS 运算符意义

运算符类型	运 算 符	运算符意义
算 术 运 算 符	+	加法
	-	减法
	*	乘法
	/	除法
	**	乘幂
	()	括号
关 系 运 算 符	=	等于
	>	大于
	<	小于
	>=	大于等于
	<=	小于等于
	~=	约等于
逻辑 运算 符	&	与
		或
	~	非

4. Function group列表框和Functions and Special Variables框

函数组列表框和函数与特殊变量框。Function group 列表框里列举了 SPSS 的所有函数组，单击任意一个组名，这一组中所有的函数和特殊变量将出现在 Functions and Special Variables 框内，再单击任意一个函数名，这个函数的信息就出现在计算器板下面的空白框中，供用户查询这个函数的意义和用法。这样的设计与 SPSS 以前版本中只用一个 Function 框列举出所有函数的方法相比，更加方便。

5. If按钮及If Cases子对话框

单击【If】按钮后，弹出【If Cases】子对话框，此对话框与 3.2.7 节中的【If】子对话框几乎完全相同，这里不再重复介绍该子对话框的具体元素了。

现在简要介绍一下这个子对话框的意义。在计算新变量时，可以根据统计分析的要求只需要对部分符合要求的观测量进行计算，这时可以通过这个对话框选择需要计算新变量的观测量，而排除不需要的那些观测量。

利用【Compute Variable】过程可以做很多事，最常见的功能就是利用原始变量生成新变量。这里举一个统计分析中非常重要的随机数生成的例子，该例综合应用了【Transform】菜单下的【Compute Variable】过程和【Random Number Generators】过程。

例 3.9 产生两列完全相同的随机数，它们的分布服从标准正态分布 $N(0, 1)$ 。具体步骤如下：（这里已先建立了一个新数据文件，其中有一个变量 id，用于标示随机数的个数。）

执行【Transform】/【Random Number Generators】命令，弹出【Random Number Generators】对话框	
勾选 Set Starting Point 选项，选择“Fixed value”	
Value : 110	设置随机种子，保证生成随机数的可重复性
执行【Transform】/【Compute Variable】命令，弹出【Compute Variable】对话框	
Target Variable : randomnum1	在目标函数框中填入 randomnum1
Function group : Random Numbers	在 Function group 框中选择 Random Numbers
Functions and Special Variables : Rv.Normal	在 Functions and Special Variables 框中选择 Rv.Normal
Numeric Expression : RV.NORMAL(0,1)	将此选项选入 Numeric Expression 栏，并输入参数 0 和 1
单击【OK】按钮	文件中出现一列随机数
重复刚才的所有步骤，只是将目标变量名改为 randomnum2	得到如图 3-30 所示的数据文件

从图 3-30 可知，在本例中，由于对两组随机数预先设定了相同的随机数种子，所以产生了两组完全相等的随机数。

	id	randomnum1	randomnum2
1	1	0.22	0.22
2	2	2.62	2.62
3	3	0.95	0.95
4	4	0.06	0.06
5	5	-1.13	-1.13
6	6	1.29	1.29
7	7	-0.63	-0.63
8	8	1.70	1.70
9	9	0.22	0.22
10	10	0.45	0.45
11	11	0.21	0.21
12	12	0.73	0.73
13	13	1.87	1.87
14	14	-0.96	-0.96
15	15	-0.31	-0.31

图 3-30 两组完全相同的服从标准正态分布的随机数

3.3.2 变量值标识——Count Values within Cases过程

变量值标识是指利用【Count Values within Cases】过程，标识某个或某几个变量值是否在观测量中出现过。

执行【Transform】/【Count Values within Cases】命令，弹出如图 3-31 所示的【Count Values within Cases】对话框，首先介绍其中的主要元素。

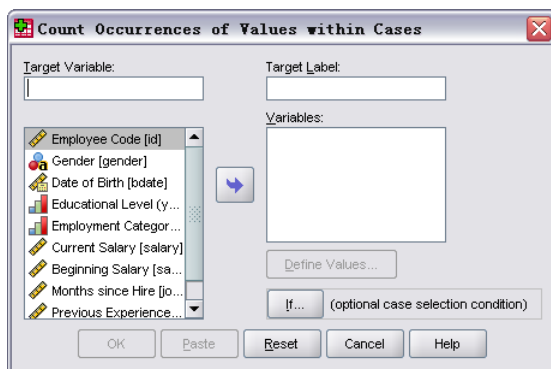


图 3-31 【Count Values within Cases】对话框

1. Target Variable栏和Target Label栏

目标变量栏和目标变量标签栏。在 Target Variable 栏内输入目标变量名，以保存标识结果；在 Target Label 栏内输入目标变量的标签。

2. Variables栏

变量栏，输入将对其进行特定变量值标识的变量。

注意 移入该栏的所有变量必须具有相同的类型。

3. Define Values按钮

单击【Define Values】按钮，弹出如图 3-32 所示的【Values to Count】子对话框，其中包含了两个框，一个是 Value 单选栏，另一个是 Values to Count 框。

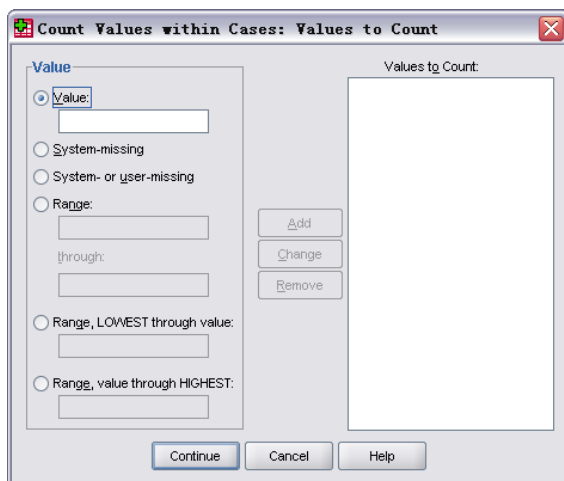


图 3-32 【Values to Count】子对话框

- Value 单选框

取值单选框，其中有 6 个单选项，其中：

Value，若观测量的变量取值等于给定的变量值，则标识该观测量，在空白栏中填入这个指定值；

System missing，若观测量的变量取值为系统缺失值，则标识该观测量；

System or user-missing，若观测量的变量取值为系统缺失值或用户自定义缺失值，则标识该观测量；

Range，若观测量的变量取值在给定的范围内，则标识该观测量，将确定这个范围的两个值分别输入 **Range** 单选项下的两个空白框中；

Range, LOWEST through value，当观测量的变量取值在最小值到指定值之间，则标识该观测量，将这个指定值输入到这个选项下的空白栏中；

Range, value through HIGHEST，当观测量的变量取值在指定值到最大值之间，则标识该观测量，将这个指定值输入到这个选项下的空白栏中。

- Values to Count 框

标识值框。在 **Value** 单选框内选定后，单击【Add】按钮，这个值或者范围就加入到 **Values to Count** 框中了。还可以通过单击【Change】按钮改变以前的选择，或者单击【Remove】撤销以前的选择。

4. If Cases子对话框

用于选择需要标识的观测量，其操作与前面类似，这里不再复述。

现举个例子来说明【Count Values within Cases】过程的应用。

例 3.10 要求在“Employee Data.sav”文件中，标识工资在 30 000 元以上，学历在 14~18 年之间的员工。标识变量名设为 s_ed，变量标签为工资学历标识。具体操作如下：

执行【Transform】/【Count Values within Cases】命令，弹出【Count Values within Cases】对话框

Target Variable : s_ed	目标变量名设为 s_ed
Target Label : 工资学历特定值标识	目标变量标签为工资学历标识
Variables : salary	将变量 salary 选入 Variables 栏
单击【Define Values】按钮	进入【Values to Count】子对话框
Value : Range, value through HIGHEST	选择 Value 单选框中最后一项
Range, value through HIGHEST : 30000	指定值设为 30000
单击【Add】按钮	将特定值选入 Values to Count 框
单击【Continue】按钮	返回【Count】对话框
Variables : educ	将变量 educ 选入 Variables 栏
单击【Define Values】按钮	进入【Values to Count】子对话框
Value : Range	选择 Value 单选框中第四项
Range : 14 through 18	指定范围为 14 到 18
单击【Add】按钮	将特定值选入 Values to Count 框
单击【Continue】按钮	返回【Count】对话框
单击【OK】按钮	新变量出现在原文件中，如图 3-33 所示

id	gender	bdate	educ	jobcat	salary	salbegin	jobtime	prevexp	minority	s_ed
1 ...		02/03/1952	15	Manager	\$57,000	\$27,000	98	144	No	2.00
2 ...		05/23/1958	16	Clerical	\$40,200	\$18,750	98	36	No	2.00
3 ...		07/26/1929	12	Clerical	\$21,450	\$12,000	98	381	No	0.00
4 ...		04/15/1947	8	Clerical	\$21,900	\$13,200	98	190	No	0.00
5 ...		02/09/1955	15	Clerical	\$45,000	\$21,000	98	138	No	2.00
6 ...		08/22/1958	15	Clerical	\$32,100	\$13,500	98	67	No	2.00
7 ...		04/26/1956	15	Clerical	\$36,000	\$18,750	98	114	No	2.00
8 ...		05/06/1966	12	Clerical	\$21,900	\$9,750	98	missing	No	0.00
9 ...		01/23/1946	15	Clerical	\$27,900	\$12,750	98	115	No	1.00
10 ...		02/13/1946	12	Clerical	\$24,000	\$13,500	98	244	No	0.00

图 3-33 经过变量值标识过后的部分新文件

如图 3-33 所示，当观测量同时满足两个条件时，其对应标识变量取值为 2；若只满足一个条件其标识变量取值为 1；否则取值为 0。

3.3.3 变量重新赋值——Recode into Same Variables/ Recode Into Different Variables过程

变量重新赋值功能是指将数据文件中的原变量值按照某种一一对应的关系生成新变量值。这里可以用这个新变量值替代原变量值，也可以生成一个新变量的过程。

执行【Transform】/【Recode into Same Variables】命令，弹出如图 3-34 所示的【Recode into Same Variables】对话框。如果选择这个命令过程，系统将用产生的新变量值直接替代原变量值。

执行【Transform】/【Recode Into Different Variables】命令，弹出如图 3-35 所示的【Recode into Different Variables】（重新赋值给不同变量）对话框，选择这个命令，系统将把产生的新变量值赋给一个新的变量。从两个图中可以看到，【Recode into Different Variables】对话框只是比【Recode into Same Variables】对话框多了一个“Output Variable”框，所以这里就以【Recode into Different Variables】对话框为例来介绍其中的主要元素。

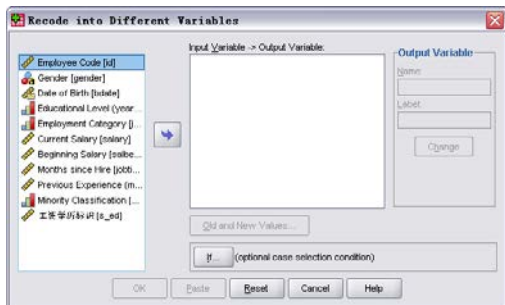
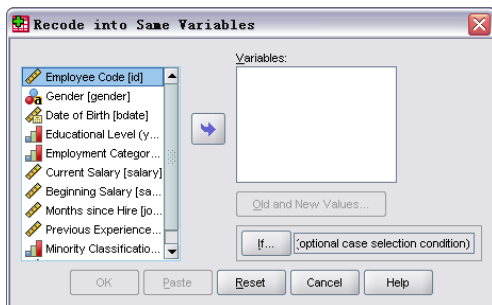


图 3-34 【Recode into Same Variables】对话框 图 3-35 【Recode into Different Variables】对话框

1. Input Variable → Output Variable框

输入变量→输出变量框。其中显示将要重新赋值的原变量名和将要建立的新变量名。

注意 可以将多个变量移入其中，但是这些变量的类型必须相同。

2. Output Variable框

输出变量框。用于定义新变量的变量名 Name 和变量标签 Label。输入完成后单击【change】按钮，新变量名出现在 Input Variable → Output Variable 框中。

3. Old and New Values子对话框

单击【Old and New Values】按钮，弹出如图 3-36 所示的【Old and New Values】子对话框。该对话框主要由 3 块组成。

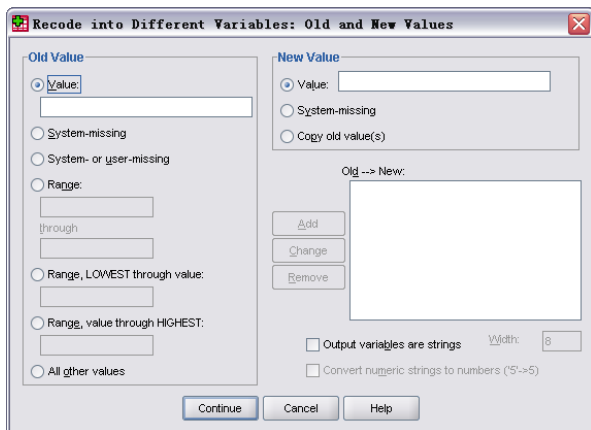


图 3-36 【Old and New Values】子对话框

- Old Value 单选框

原变量值框，用于选择将赋新值的变量值。类似于【Count】过程中【Values to Count】子对话框的 Value 框，只比它多了一项 All other values（所有其他变量值）选项，其他的选项不再一一介绍。

- New Value 单选框

新变量值框，用于选择将赋予的新变量值，其中有三个选项：

Value，直接为新变量赋予一个新的指定值，这个指定值就填入 Value 选项后的空白栏内；

System missing，为新变量赋予系统缺失值；

Copy Old Value，将原变量值直接赋给新变量值。

- Old → New 框

原变量值→新变量值框，显示由原变量值转化为新变量值的详细信息。在选定 Old Value 单选框中选项和 New Value 单选框中选项后，单击【Add】按钮，信息就显示在这个框中。

- Output Variables are strings 选项

新变量值赋予字符型变量选项。选择此项，无论原变量值是否为字符型都将被赋值为字符型变量。

- Converts numeric string to numbers（‘5’→5）选项

将以数值作为字符串的字符型变量转换为数值型变量。只有选入的原变量为字符型变量时，此项才会被激活。

4. If Cases子对话框

用于选择观测量。

下面举例说明这一过程的用法。

例 3.11 要求在“Employee Data.sav”文件中，将原变量 educ 中的值教育的年数重新赋值为新变量 edulever（教育的等级），系统缺失值仍为系统缺失值，教育年数为 8 的转化为第 1 等级，教育年数为 12~15 之间的转化为第 2 等级，教育年数为 16~19 的转化为第 3 等级，教育年数为 19 及以上的转化为第 4 等级。具体操作如下：

执行【Transform】/【Recode Into Different Variables】命令，弹出【Recode Into Different Variables】对话框			
Input Variable	Output Variable 框：educ	将变量 educ 选入这个框中	
Output Variable：			
Name：edulever		定义新变量名为 edulever	
Label：教育等级		定义变量标签为教育等级	
单击【Change】按钮		添加新变量名	
单击【Old and New Values】按钮		弹出【Old and New Values】子对话框	
Old	New:SYSMIS	SYSMIS	按要求选择好转化的方法
	8	1	
	12 thru 15	2	
	16 thru 18	3	
	19 thru Highest	4	

单击【Continue】按钮

回到主对话框

单击【OK】按钮

新变量出现在文件中，如图 3-37 所示

id	gender	bdate	educ	jobcat	salary	salbegin	jobtime	prevexp	minority	edulevel
1	Male	02/03/1952	15	Manager	\$57,000	\$27,000	98	144	No	2.00
2	Male	05/23/1958	16	Clerical	\$40,200	\$18,750	98	36	No	3.00
3	Female	07/26/1929	12	Clerical	\$21,450	\$12,000	98	381	No	2.00
4	Female	04/15/1947	8	Clerical	\$21,900	\$13,200	98	190	No	1.00
5	Male	02/09/1955	15	Clerical	\$45,000	\$21,000	98	138	No	2.00
6	Male	08/22/1958	15	Clerical	\$32,100	\$13,500	98	67	No	2.00
7	Male	04/26/1956	15	Clerical	\$36,000	\$18,750	98	114	No	2.00
8	Female	05/06/1966	12	Clerical	\$21,900	\$9,750	98	missing	No	2.00
9	Female	01/23/1946	15	Clerical	\$27,900	\$12,750	98	115	No	2.00
10	Female	02/13/1946	12	Clerical	\$24,000	\$13,500	98	244	No	2.00
11	Female	02/07/1950	16	Clerical	\$30,300	\$16,500	98	143	No	3.00
12	Male	01/11/1966	8	Clerical	\$28,350	\$12,000	98	26	Yes	1.00

图 3-37 重新赋值后的部分数据文件

注意 【Record】过程一个重要应用即是从原始数据生成简易频数表。请读者自己思考一下如何利用【Record】过程实现这一功能。具体实例可参见本书第六章例 6.1。

3.3.4 变量值秩排序——Rank Cases过程

上一节讲过观测量排序，【Rank Cases】过程是一项更加强大的排序功能。它不仅能对观测量按照给定变量取值进行排序，还能将排序结果赋予各个观测量，即通常所说的求秩。

执行【Transform】/【Rank Cases】命令，弹出如图 3-38 所示的【Rank Cases】对话框，首先介绍其中主要元素。

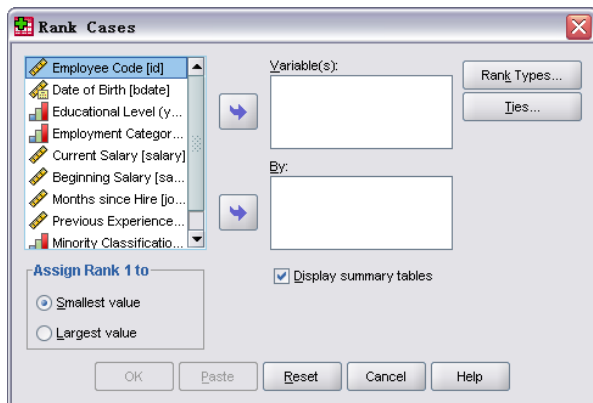


图 3-38 【Rank Cases】对话框

1. Assign Rank 1 to 单选框

选择将秩 1 赋给最大值还是最小值。其中有两个选项：

- **Smallest Value**，将秩 1 赋给最小变量值，即从最小值开始按照升序对观测量排序；
- **Largest Value**，将秩 1 赋给最大变量值，即从最大值开始按照降序对观测量排序。

2. Variable(s)栏

变量栏，以该栏中变量的取值作为观测量排序以及求秩变量的依据。新生成的秩变量名就是原始变量名字前加字母“r”。

3. By变量栏

分组变量栏，按照该变量将原始观测量分成多个小组，在各组内分别排序和求秩。如果该变量值缺失的话，则对所有观测量一起排序和求秩。

4. Types子对话框

单击【Rank Types】按钮，弹出如图 3-39 所示的【Types】（类型）子对话框，这个对话框主要是用来选择秩的类型。

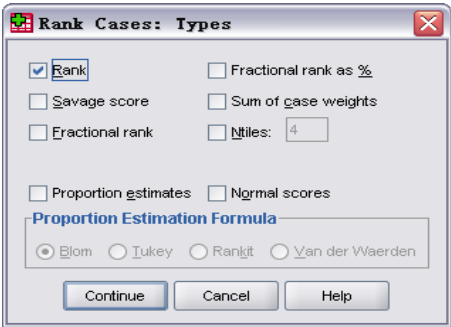


图 3-39 【Types】子对话框

各类秩的计算方法如表 3-4 所示。

表 3-4 SPSS 秩类型

秩		类 型
Rank		基本秩，此为系统默认值
Savage score		秩变量取值为根据指数分布得到的 Savage 得分
Fractional rank		小数秩，秩变量值为秩除以非缺失值权重之和的商
Fractional rank as %		百分数秩，秩变量值为秩除以非缺失值权重之和的商的百分数形式
Sum of case weights		权重和，秩变量值为各观测量值权重之和。对同一分组中的所有观测量，秩变量取值为一个常数
Ntiles		对观测值作百分位数分组，各组中观测值的秩分别为所在组的组序号，分组的数目由此栏后面的空白栏中的数确定，因此空白栏内要输入一个大于 1 的整数
Normal scores		正态得分
Proportion estimates (比例估计)	Blom	公式为 $(r-3/8) / (w+1/4)$
	Tukey	公式为 $(r-1/3) / (w+1/3)$
	Rankit	公式为 $(r-1/2) / w$
	Van der Waerden	公式为 $r / (w+1)$

说明：（1）公式中字母“w”指观测量权重的和；（2）公式中字母“r”指秩

5. Ties子对话框

单击【Ties】按钮，弹出如图 3-40 所示的【Ties】子对话框，这个对话框主要是用来处理相等观测量的秩问题。

其中各选项的含义如下：

- **Mean**，取各个相等观测量秩的均值为相同观测量处的秩，此为系统默认值；
- **Low**，取各个相等观测量秩的最小值为相同观测量处的秩；
- **High**，取各个相等观测量秩的最大值为相同观测量处的秩；
- **Sequential ranks to unique values**，把相同的观测量当作一条记录处理。

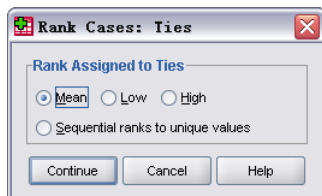


图 3-40 【Ties】子对话框

6. Display summary table选项

显示摘要信息表选项，选择它则系统将在输出窗口中显示概括原变量和新变量的摘要信息表。系统默认是显示。

现举一个例子来说明【Rank Cases】过程的应用

例 3.12 要求在“Employee Data.sav”文件中，对所有雇员按照受教育年限分组，在各组中按照当前工资水平进行排序，并生成其相应的秩变量。具体操作步骤如下：

执行【Transform】/【Rank Cases】命令，弹出【Rank Cases】对话框

Variable : Salary

定义工资为排序变量

By : Educ

定义受教育年限为分组变量

单击【OK】按钮

新变量出现在文件中，如图 3-41 所示

	id	gender	bdate	educ	jobcat	salary	salbegin	jobtime	prevsup	minority	Rsalary
1	1	Male	02/03/1952	15	Manager	\$57,000	\$27,000	98	144	No	115,000
2	2	Male	05/23/1958	16	Clerical	\$40,200	\$18,750	98	36	No	21,500
3	3	Female	07/26/1929	12	Clerical	\$21,450	\$12,000	98	381	No	35,500
4	4	Female	04/15/1947	8	Clerical	\$21,900	\$13,200	98	190	No	20,500
5	5	Male	02/09/1955	15	Clerical	\$45,000	\$21,000	98	138	No	109,000
6	6	Male	08/22/1958	15	Clerical	\$32,100	\$13,500	98	67	No	75,000
7	7	Male	04/26/1956	15	Clerical	\$36,000	\$18,750	98	114	No	93,000
8	8	Female	05/06/1966	12	Clerical	\$21,900	\$9,750	98	missing	No	40,500
9	9	Female	01/23/1946	15	Clerical	\$27,900	\$12,750	98	115	No	43,000
10	10	Female	02/13/1946	12	Clerical	\$24,000	\$13,500	98	244	No	75,000
11	11	Female	03/07/1950	16	Clerical	\$30,300	\$15,500	98	143	No	5,000
12	12	Male	01/11/1966	8	Clerical	\$29,350	\$12,000	98	26	Yes	36,000
13	13	Male	07/17/1960	15	Clerical	\$27,750	\$14,250	98	34	Yes	40,500
14	14	Female	02/26/1949	15	Clerical	\$35,100	\$16,800	98	137	Yes	88,500
15	15	Male	08/29/1962	12	Clerical	\$27,300	\$13,500	97	66	No	125,000
16	16	Male	11/17/1964	12	Clerical	\$40,800	\$15,000	97	24	No	187,500
17	17	Male	07/18/1962	15	Clerical	\$46,000	\$14,250	97	48	No	111,000
18	18	Male	03/20/1956	16	Manager	\$103,750	\$27,510	97	70	No	59,000
19	19	Male	08/19/1962	12	Clerical	\$42,300	\$14,250	97	103	No	189,000
20	20	Female	01/23/1940	12	Clerical	\$26,250	\$11,550	97	48	No	110,000
21	21	Female	02/19/1963	16	Clerical	\$38,850	\$15,000	97	17	No	20,000
22	22	Male	09/24/1940	12	Clerical	\$21,750	\$12,750	97	315	Yes	36,500
23	23	Female	03/15/1965	15	Clerical	\$24,000	\$11,100	97	75	Yes	13,000
24	24	Female	03/27/1933	12	Clerical	\$16,950	\$9,000	97	124	Yes	5,500
25	25	Female	07/01/1942	15	Clerical	\$21,150	\$9,000	97	171	Yes	4,000

图 3-41 秩排序之后数据文件

如图 3-41 所示，新生成的数据文件中包含了一系列新增的秩变量“Rsalary”，该变量用来记录各个雇员薪酬水平在相同受教育年限组中的秩排序。

3.3.5 Transform菜单其他过程简介

上面介绍了 Transform 菜单中几个非常重要的数据文件整理的功能。剩下的一些特殊功能简要介绍如表 3-5 所示。

表 3-5 Transform 菜单其他命令简介

名 称	作 用
【Shift Values】	转换变量值。生成一个新变量，其取值为原始变量向上或向下平移 N 个单位的取值
【Automatic Recode】	自动重新赋值功能。按照原变量大小生成新变量，新变量取值就是原变量的大小次序
【Visual Binning】	可视化离散。依据某一个或多个变量取值将观测量离散化
【Optimal Binning】	最优离散化，根据选择变量取值自动离散化观测量
【Data and Time Wizard】	日期时间向导。用于简化很多关于日期和时间变量的共同操作
【Create Time Series】	自动生成时间序列变量，主要用于时间序列模型
【Replace Missing Values】	缺失值的替代处理，当序列中有缺失值时，采用方法替代缺失值，并将结果记入一个新变量中
【Random Number Generators】	随机数生成器。用于选择随机数生成器以及设置随机数生成的开始点（又叫种子）。利用同一个开始点可以产生两列完全相同的随机数，在例 3-9 中见到过它的用法

3.4 本章小结

本章介绍了 SPSS 中【Data】菜单和【Transform】菜单的多个过程，其中【Data】菜单详细介绍了 8 个过程，【Transform】菜单详细介绍了 4 个过程。在这些过程中，最常用的过程如下所示：

- 【Data】菜单

观测量排序——Sort Case 过程；

数据文件的拆分——Split File 过程；

选择观测量——Select Cases 过程；

观测量加权——Weight Cases 过程。

- 【Transform】菜单

变量计算——Compute Variable 过程；

变量重新赋值——Recode into Same Variables/ Recode Into Different Variables 过程。

当然，对于那些这里没有提到的但是本章也详细介绍的过程，也是非常重要的。只是它们在处理实际问题的时候，没有以上提到的几个过程使用频率高。

第 4 章 SPSS统计图形

上一章介绍了数据的整理。在一个完整的统计分析过程中，整理数据作为第一步是为了方便用户使用数据，接下来用户要考虑的问题是应该用哪种统计方法来处理数据。通常，这主要根据一个科研工作者的实际经验来决定。但是，无论是一个有经验的统计高手还是只是偶尔进行统计分析的“新人”，如果在分析的过程中采用统计图形加以辅助都会达到事半功倍的效果。本章主要介绍统计分析过程中重要的工具——统计图形。本章内容包括：

- 统计图形概述
- 常见统计图形
- SPSS 图形编辑
- 交互式统计图形

4.1 统计图形概述

统计图形是数据最直观表示。科研工作者在处理实际问题时，面对大量的数据，有可能不知从何入手。这时统计图形的优越性就显示出来了。通过图形，用户可以对数据的基本特征有一个感性的认识，为进一步选取适当的统计方法和模型打下基础。同时，统计工作的服务对象多是非专业的统计人士。非专业人士往往不明白什么是相关分析、什么是回归分析等。与其大费口舌地给他们解释这些方法，不如直接用统计图形来给出直观的结果，更方便其理解。

同 SAS、S-PLUS 等统计软件相比，SPSS 的绘图一般情况下无需编程，而是由界面操作直接完成，这使初学者更易掌握。SPSS 图形美观，并能通过图形编辑窗口进一步修饰。因此，SPSS 强大的绘图功能是其与其他统计软件竞争时的一把利剑。

本节简要介绍 SPSS 的【Graphs】菜单和常用统计图形的基本特点，为后面章节的铺开叙述打下基础。

4.1.1 Graphs菜单简介

如图 4-1 所示，SPSS 的【Graphs】菜单主要由图形生成器【Chart Builder】、图形模板选择器【Graphboard Template Chooser】以及统计图形【Legacy Dialogs】构成。其中统计图形【Legacy Dialogs】又包括了多种常用统计图形和交互式图形【Interactive】，将分别在

本章 4.2、4.4 节介绍。本节首先简单介绍 SPSS 的图形生成器【Chart Builder】过程和图形模板选择器【Graphboard Template Chooser】过程。

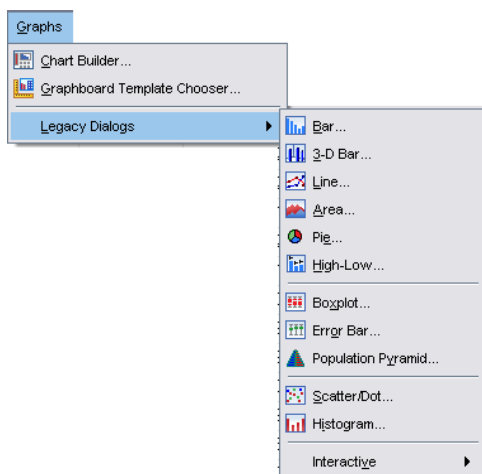


图 4-1 【Graphs】菜单

1. 图形生成器 Chart Builder

图形生成器是一个绘制特色图形的辅助工具。执行【Graphs】/【Chart Builder】命令，弹出如图 4-2 所示【Chart Builder】对话框。

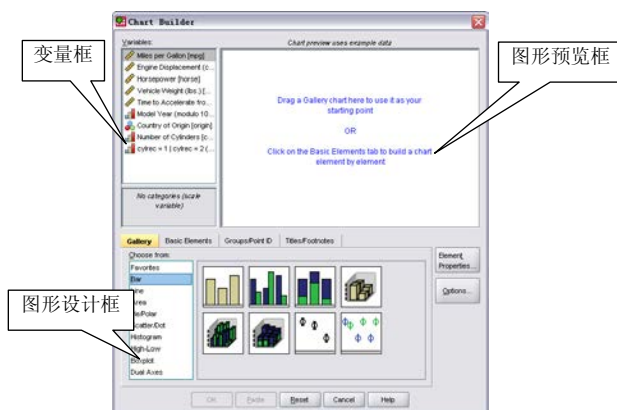


图 4-2 【Chart Builder】对话框

【Chart Builder】对话框共分为 3 部分。

- Variables 框，变量框，放置打开数据文件中的变量；
- Chart preview uses example data 框，图形预览框，在图形设计过程中用来预览定义的图形；
- 图形设计框：如图 4-3 所示，图形设计框包含 4 个选项卡，分别用来定义图形类别（Gallery）、图形中基本元素（Basic Elements）、图形分组变量（Groups/Point ID）和图形标题和脚注（Titles/Footnotes）。



图 4-3 图形设计框基本元素


注意 在使用图形生成器或交互式绘图之前必须准确定义变量的测量尺度。

下面通过一个简单的例子介绍图形生成器的使用。

例 4.1 已知某地在若干个时间点的住宅成交数据“house price.sav”如表 4-1 所示，利用图形生成器，绘制该地住房成交情况的双轴图。

表 4-1 某地住宅成交数据

时 间	住宅成交均价 (元/平方米)	住宅成交量 (万平方米)
1	5000	120
2	5100	95
3	5080	123
4	5200	115
5	5350	110
6	5530	100
7	6000	105
8	6200	130

用户首先在图形设计框选择要绘制的图形形状，切换到【Gallery】选项卡，选择双轴图（Dual Axes）。将图标拖曳至图形预览窗口。

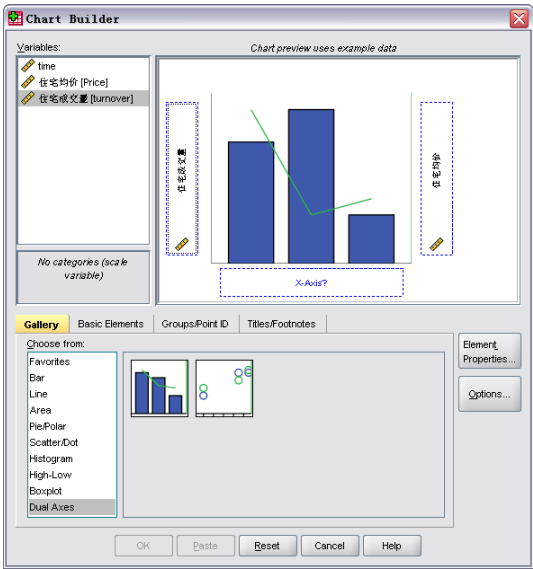


图 4-4 预设后的【图形生成器】对话框

同时，将变量“time”拖曳至 X 轴位置，变量“turnover”和“price”分别拖曳至左右 Y 轴位置。此时图形预览窗口如图 4-4 所示，单击【OK】按钮。此时，在 SPSS 浏览窗口就生成了相应的住宅成交情况双轴图。经过简单的图形编辑，可得到如图 4-5 所示的图形。

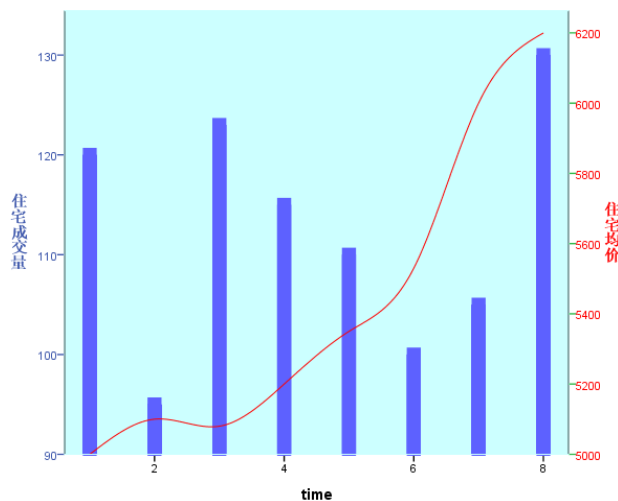


图 4-5 图形生成器绘制的双轴图

从例 4.1 可以发现，图形生成器可以交互式地绘制图形。例 4.1 只是简单地介绍了图形生成器的使用步骤，如何用图形生成器来绘制用户喜爱的图形就有待用户自己慢慢琢磨了。

2. 图形模板选择器 Graphboard Template Chooser

图形模板选择器是一项新增的功能。它可以根据用户选择的变量个数和变量类型，自动给出一些图像模板供用户选择。例如，对于例 4.1 中的数据，执行【Graphs】/【Graphboard Template Chooser】命令，弹出如图 4-6 所示的【Graphboard Template Chooser】对话框。

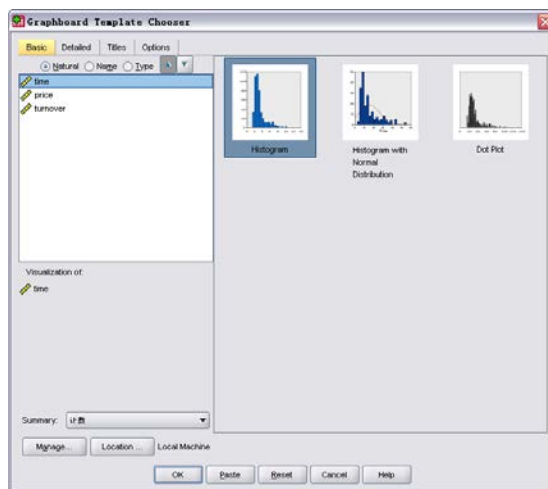


图 4-6 【Graphboard Template Chooser】对话框（一）

在图 4-6 中，仅在对话框左侧的变量框选中一个变量时，给出了 3 个图形模板备选。如果此时同时在左侧的变量框中选中两个变量，则该对话框变成如图 4-7 所示。此时备选图形模板就增加到 11 个。

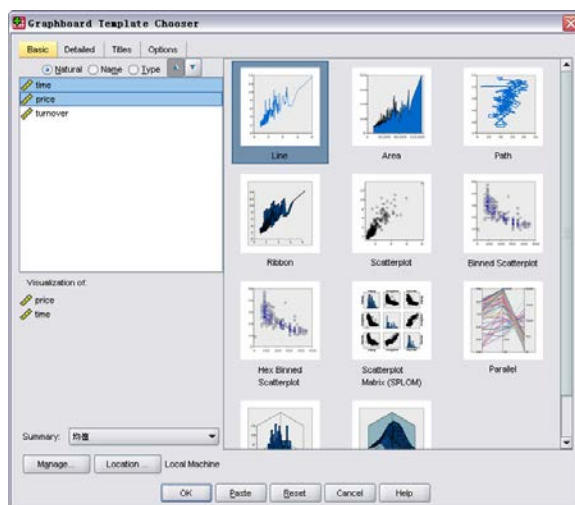






图 4-7 【Graphboard Template Chooser】对话框（二）

可见，当用户有了数据之后，却不知道怎样绘制更好的图形时，那么通过【Graphboard Template Chooser】过程必能获得很多启发式的收获。这大大方便了用户绘制统计图形。

4.1.2 常用统计图形简介

由图 4-1 可知，SPSS 的统计图形可谓包罗万象。非专业统计人员进行绘图时，面对多种多样的图形往往不知道到底应该选取哪一种更合适。因此，在具体介绍各类图形绘制方法之前，首先由表 4-2 给出常见统计图形的主要适用问题。

表 4-2 常见统计图形的适用问题

图 形 名	符 号	适用问题
条形图 (Bar Charts)		描述定类或定序变量的分布。用长条的高度来表示变量在不同取值下的频数
线图 (Line Charts)		描述连续性变量的变化趋势，非连续性变量通常不宜采用
面积图 (Area Charts)		描述连续性变量的分布。用面积来表示变量在不同取值下的频数
饼图 (Pie Charts)		描述定类变量的分布。用圆中扇形的面积大小表示不同类别变量所占的频数
高低图 (High-Low Charts)		用于同时描述股票、商品价格等市场数据长期和短期变化趋势
帕累托图 (Pareto Charts)		描述生产控制过程中各类指标对生产的影响大小

续表

图 形 名	符 号	适用问题
质量控制图（Control Charts）		质量控制的常用工具,主要用于提示生产过程中发生的变化和趋势
箱图（Boxplots）		显示变量的中位数、四分位数、极值,显示数据的实际分布
误差条图（Error Bar Charts）		显示数据的均值、标准差、置信区间等信息
散点图（Scatterplots）		直观反映两个或两个以上变量的取值大小及相互关系
直方图（Histogram）		描述定距变量的分布。与条形图不同的是直方图不是用长条的高度来表示变量出现的频数,而是通过长条的面积来表示的
P-P 图（P-P plots）		用来直观表示数据是否服从特定分布
Q-Q 图（Q-Q plots）		用来直观表示数据是否服从特定分布
普通序列图（Sequence Chart）		描述一组或几组数据随另一序列性变量变化的趋势
时间序列图（Time Series Charts）		描述与时间相关的变量随着时间变化的趋势

4.2 常见统计图形

上一节介绍了各类统计图形的主要适用情况,本节将通过具体的例子介绍各类统计图形最原始的绘制方法。对图形的编辑将在 4.3 节介绍。

4.2.1 条形图（Bar Charts）

条形图主要用于描述定类或定序变量的分布。在 SPSS 中,条形图可以分为 3 类:简单条图、分组条图和分段条图。现在分别通过 3 个例子来介绍这 3 类条图。

1. 简单条图

例 4.2 学生成绩的简单条图。已知有数据文件“chengji_1.sav”,其数据结构如表 4-3 所示,按班级绘制学生语文平均成绩的条图。

表 4-3 学生成绩统计表

姓 名	班 级	语 文	数 学	英 语
李芳	1	84.00	89.00	76.00
王明	1	76.00	93.00	89.00
张刚	1	89.00	65.00	93.00
李兰	2	93.00	82.00	65.00
.....

下面通过这个例子介绍简单条图的绘制。

执行【Graphs】/【Legacy Dialogs】/【Bar】命令，弹出如图 4-8 所示对话框。该对话框主要用于选择绘制的条图类型和定义图形中的数据。

- **【Simple】**：简单条图。
- **【Clustered】**：分组条图。
- **【Stacked】**：分段条图。
- **【Data in Chart Are】**：定义图形中数据的描述方式。从上到下共 3 类，依次为：
 观测量分类概述，主要对应于简单条图；
 变量分类概述，主要对应于分组条图；
 单个观测量值概述。

在本例中，由于要求以班级为单位绘制简单条图，所以选择**【Simple】**及“Summaries for groups of cases”，单击**【Define】**按钮，弹出如图 4-9 所示对话框。

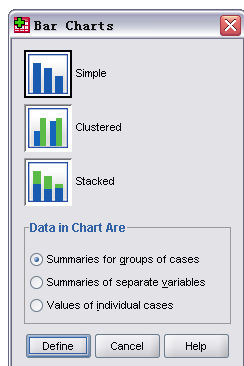


图 4-8 【Bar Charts】对话框

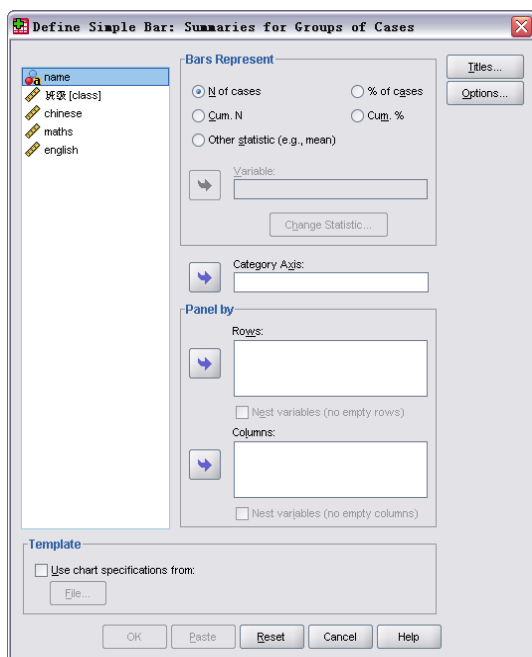


图 4-9 简单条图定义对话框

在图 4-9 所示的对话框中，最左边为变量框。请读者仔细观察变量“class”。在数据文件中，定义了变量名“class”的变量标签为“班级”。而此时该变量直接显示的是变量标签的值，变量名反而放到括号中了。在绘制图形的时候，图形上也只会显示变量标签的值。

注意 在绘制图形之前，除了要准确定义变量的测量尺度，变量标签、变量值标签也是很重要的。

下面依次介绍图 4-9 中的各选项。

- **【Bars Represent】**：定义条图中长条的具体含义。

- ① N of cases: 长条代表记录个数。
- ② % of cases: 长条代表记录的百分比。
- ③ Cum.N: 长条代表从前到后的累计记录数。
- ④ Cum.%: 长条代表从前到后的累计百分比。

⑤ Other statistic: 长条代表给定变量的某个统计值。此时，首先要选入一个变量到【Variable】框。系统默认长条代表的是该变量的均值。如果用户想自己设定的话，单击【Change Statistic】按钮，弹出如图 4-10 所示对话框。在对话框中，用户可以定义长条的含义。

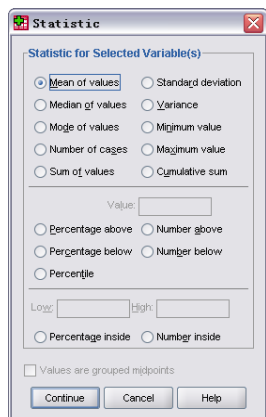


图 4-10 从上到下共 3 部分，下面依次讲解：

- 如果选取“Statistic for Selected Variables”单选框组中的任何一个，则长条代表刚才选择变量的均值、方差、中位数、极值等。
- 如果选中第二个单选框组中的任何一个，则长条代表在满足给定条件的记录个数。比如，给定“Value”为“80”，选择“Percentage above”，则长条表示组中变量值大于 80 的记录数占该组记录总数的比例。
- 如果选择的是“Number above”则长条表示组中变量值大于 80 的记录个数。若选择“Percentile”，则长条代表 80%分位点的变量值。

第三个单选框组和第二组的意义基本一致。只是第二组的选择条件是一个界限，而第三组的选择条件是一个范围。

在本例中，由于长条表示学生语文平均成绩，所以在图 4-9 所示的对话框中选择“Other statistic”，将变量“chinese”移到【Variable】框。此时，系统默认长条代表变量“chinese”的均值。

• 【Category Axis】

分类轴，代表条图的横坐标，即绘制条图时的分类变量。在本例中，将变量“班级”移入。

• 【Panel by】

图组变量框，分为行、列两个选项。用于选择变量，按照变量取值不同在同一坐标轴内绘制多张条图。

• 【Template】

选择绘制图形的模板。图形模板的定义将在 4.3 节图形编辑部分介绍。

• 【Titles】

单击【Titles】按钮，弹出如图 4-11 所示的对话框，设置图形的主标题、副标题和脚注。

• 【Options】

单击【Options】按钮，弹出如图 4-12 所示对话框，用来定义与缺失值有关的选项。

- ① Missing Values: 定义对缺失值的处理方式。
- ② Display groups defined by missing values: 选择是否把分类变量的缺失值作为一个组表示出来。
- ③ Display chart with case labels: 选择是否把变量值在图中显示为相应点的标签。只有在图中有散点且变量标签存在时此项才可用。

④ Error Bars Represent: 设置图形的置信度、标准差，等等。



图 4-11 【Titles】对话框

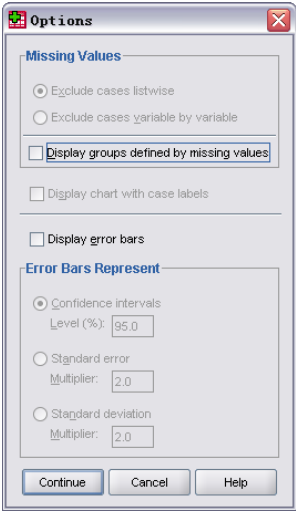


图 4-12 【Options】对话框

以上介绍了简单条图对话框的定义，现将本例的操作过程小结如下：

执行【Graphs】/【Legacy Dialogs】/【Bar】命令，弹出【Bar Charts】对话框	
选择“simple”、“Summaries for groups of cases”	选择绘制简单条图
单击【Define】按钮	弹出简单条图定义对话框
【Bars Represent】：选择 Other statistic	选择自定义条图中长条的含义
【Variable】：chinese	定义长条代表语文平均成绩
【Category Axis】：class	选择变量“class”作为分类变量
单击【Titles】按钮	定义图形的标题、脚注等
单击【OK】按钮	定义完成

执行以上过程，弹出如图 4-13 所示条图，这就是本例所绘制的未经编辑的简单条图。

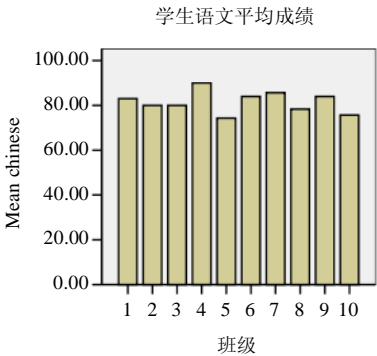


图 4-13 简单条图

2. 分组条图

分组条图是指每组有两个及两个以上长条组成的图形。下面仍通过例子来介绍其用法。

例 4.3 学生成绩的分组条图。已知有数据文件“chengji_2.sav”，其数据结构如表 4-4 所示。已知共有 5 个年级，每个年级有 3 个班。

- (1) 以年级分组绘制各班语文平均成绩的条图。
- (2) 绘制各年级语文、数学、英语平均成绩的条图。

表 4-4 学生成绩统计表

姓 名	年 级	班 级	语 文	数 学	英 语
李芳	1	1	84.00	89.00	76.00
王明	1	2	76.00	93.00	89.00
张刚	1	3	89.00	65.00	93.00
李兰	2	1	93.00	82.00	65.00
.....

本例的两个问题正好代表了绘制两种不同分组条图的方法。

对于问题（1）执行如下操作：

执行【Graphs】/【Legacy Dialogs】/【Bar】命令，弹出【Bar Charts】对话框	
选择“cluster”、“Summaries for groups of cases”	选择绘制分组条图 1，弹出如图 4-14 所示对话框
【Bars Represent】：选择 Other statistic	选择自定义条图中长条的含义
【Variable】：chinese	定义长条代表语文平均成绩
【Category Axis】：grade	选择变量“grade”作为分类变量
【Define Clustered by】：class	选择变量“class”作为分组变量
单击【Titles】按钮	定义图形的标题、脚注等
单击【OK】按钮	定义完成

生成如图 4-15 所示的分组条图。

对于问题（2）执行如下操作：

执行【Graphs】/【Legacy Dialogs】/【Bar】命令，弹出【Bar Charts】对话框	
选择“cluster”、“Summaries of separate variables”	选择绘制分组条图 2，弹出如图 4-16 所示对话框
【Bars Represent】：chinese、maths、english	定义每类中显示的三组信息
【Category Axis】：grade	选择变量“grade”作为分类变量
单击【Titles】按钮	定义图形的标题、脚注等
单击【OK】按钮	定义完成

生成如图 4-17 所示分组条图。

比较图 4-14 和图 4-16 可以发现，虽然二者都是定义分组条图。但是，有个本质的区别就是在第 1 类分组条图中长条代表的含义是一样的，都是语文成绩。不同的长条代表同

一年级中的不同班级。而第2类分组条图中长条代表的含义是不同的，它分别代表着同一年级中语、数、外三门课程的平均成绩。

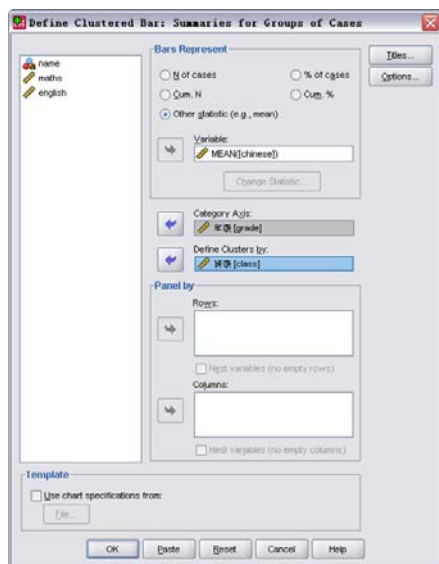


图 4-14 第1类分组条图定义框

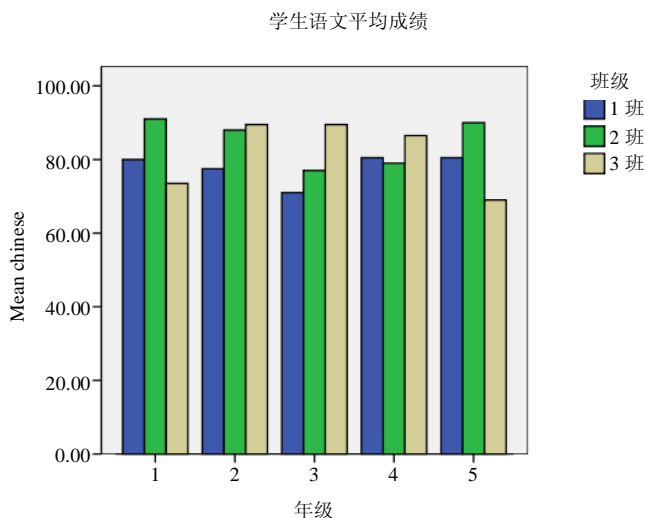


图 4-15 例 4.3 (1) 生成的分组条图

图 4-15 和图 4-17 分别对应着例 4.3 中的两个问题。虽然二者同为分组条图，从图形上乍一看也有点相近。但是仔细比较二者的图标说明就可以发现它们的区别了。

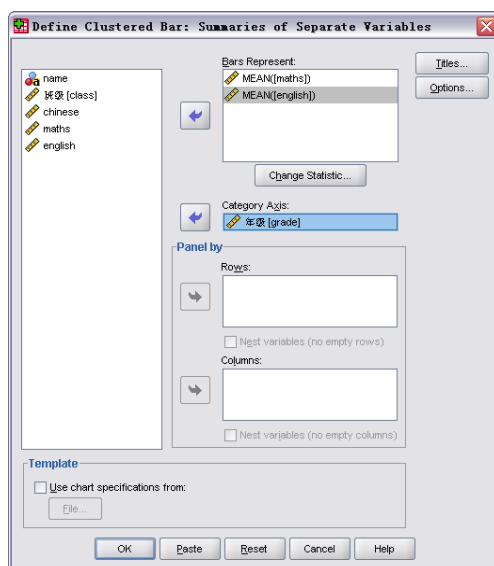


图 4-16 第2类分组条图定义框

3. 分段条图

分段条图以长条的整体代表变量的整体，长条的各段代表各组在整体中所占的比例。可以说分段条图是将分组条图中的各组信息叠加到一个长条上来的。当然，二者在统计描

述上还是有不同的侧重点的。分组条图更侧重于刻画各组之间的比较关系。而分段条图则侧重于刻画各段与整体之间的关系。

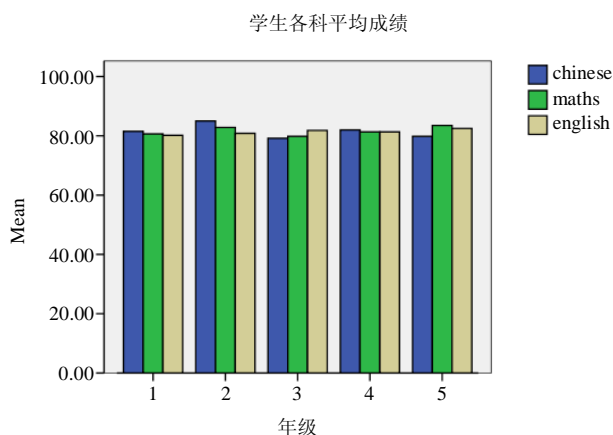


图 4-17 例 4.3 (2) 生成的分组条图

例 4.4 学生成绩的分段条图。仍然采用例 4.3 中的数据文件“chengji_2.sav”，绘制各年级语、数、外平均总成绩的分段条图。

执行如下操作：

执行【Graphs】/【Legacy Dialogs】/【Bar】命令，弹出【Bar Charts】对话框	
选择“stacked”、“Summaries of separate variables”	选择绘制分段条图
【Bars Represent】: chinese、maths、english	定义每类中显示的三组信息
【Category Axis】: grade	选择变量“grade”作为分类变量
单击【Titles】按钮	定义图形的标题、脚注等
单击【OK】按钮	定义完成

生成如图 4-18 所示的分段条图。

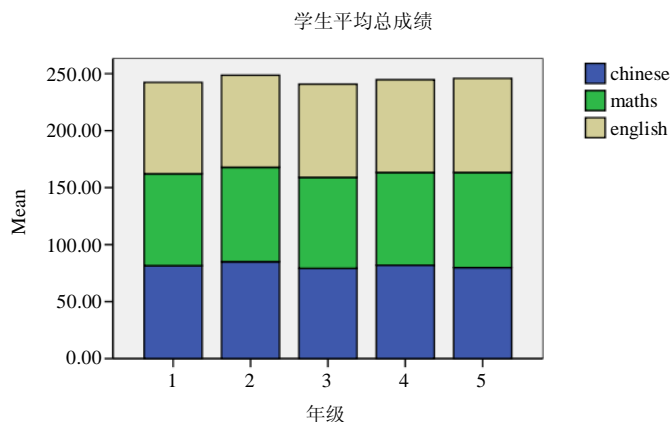


图 4-18 分段条图

前面所介绍的一般条图都只有一个分类变量，3-D 条图相当于在一般条图的基础上再引入了一个分类变量，其他地方和一般条图类似。

4.2.2 线图（Line Charts）

SPSS 的线图绘制是与条图完全对应的。执行【Graphs】/【Legacy Dialogs】/【Line】命令，弹出如图 4-19 所示对话框。

比较线图对话框图 4-19 和条图对话框图 4-8 可以发现，二者非常相似，都是由选择图形类型和定义图形中数据两部分组成的。其实不仅仅对于条图和线图，在 SPSS 中几乎所有的图形的第一个对话框都是这样的。

线图分为简单线图、复式线图和垂线图 3 类。通过比较可以发现，简单线图与简单条图的定义对话框类似，复式线图与分组条图的定义对话框一样。它们完全是相互对应的关系。

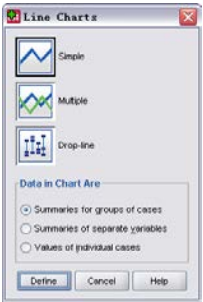


图 4-19 【Line Charts】对话框

例 4.5 学生成绩的线图。仍以数据文件“chengji_1.sav”为例，绘制如下线图：

- (1) 绘制各班语文平均成绩的简单线图。
- (2) 绘制各班三门成绩平均分的复式线图。

对于问题（1）执行如下操作：

执行【Graphs】/【Legacy Dialogs】/【Line】命令，弹出【Line Charts】对话框	
选择“simple”、“Summaries for groups of cases”	选择绘制简单线图
【Line Represent】：选择 Other statistic	选择自定义线图中线条的含义
【Variable】：chinese	定义线条代表语文平均成绩
【Category Axis】：class	选择变量“class”作为分类变量
单击【Titles】按钮	定义图形的标题、脚注等
单击【OK】按钮	定义完成

生成如图 4-20 所示的简单线图。

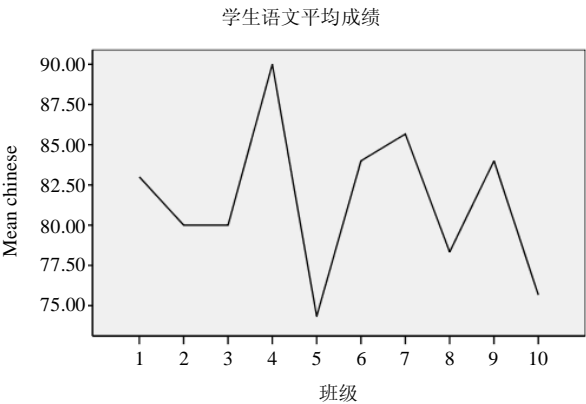


图 4-20 简单线图

对于问题（2）执行如下操作：

执行【Graphs】/【Legacy Dialogs】/【Line】命令，弹出【Line Charts】对话框	
选择“multiple”、“Summaries of separate variables”	选择绘制复式线图
【Line Represent】：chinese、maths、english	定义每类中显示的三组信息
【Category Axis】：class	选择变量“class”作为分类变量
单击【Titles】按钮	定义图形的标题、脚注等
单击【OK】按钮	定义完成

生成如图 4-21 所示的复式线图，其中一条直线代表了一门学科的成绩信息。

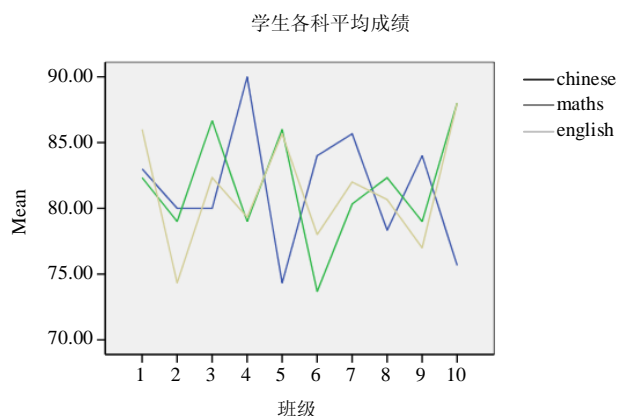


图 4-21 复式线图

垂线图是指用同一条直线上点的高低来代表变量取值的大小。利用数据文件“chengji_1.sav”绘制的垂线图如图 4-22 所示。

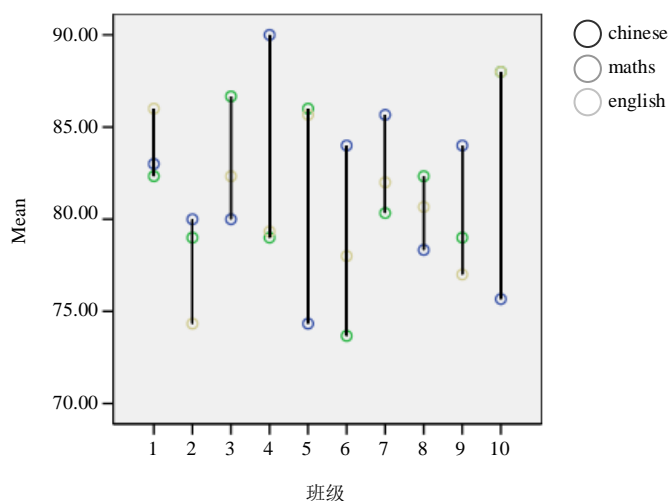


图 4-22 垂线图

线图和条图只是采用了不同表现形式的两种统计图形。

4.2.3 面积图 (Area Charts)

条图、线图和面积图三者都是用来描述变量的分布情况，并且可以相互转换。面积图的定义同前面二者类似，这里也就不再详述了。

与条图和线图相比，面积图唯一的区别是面积图只分为简单面积图和分段面积图两种。简单面积图与简单条图、简单线图相对应，分段面积图与分段条图相对应。图 4-23 是利用数据文件“chengji_1.sav”绘制的分段面积图。比较分段面积图和分段条图可以发现，我们可以将分段面积图看做是由离散的分段条图连续化得出的新图形。

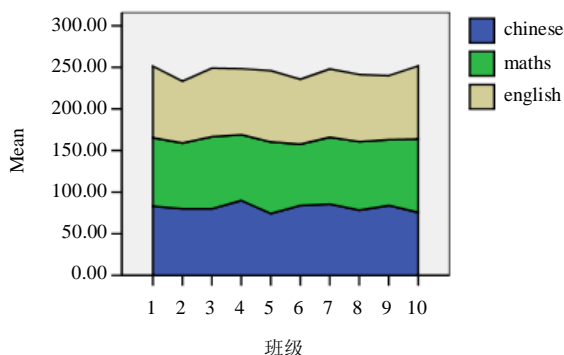


图 4-23 分段面积图

4.2.4 饼图 (Pie Charts)

线图、条图和面积图都是描述变量在不同取值下的分布。饼图则是用来表示部分与整体之间的关系。

例 4.6 某店营业额。已知有数据文件“chaoshi.sav”，绘制某店营业额的饼图。

执行如下操作：

执行【Graphs】/【Legacy Dialogs】/【Pie】命令，弹出【Pie Charts】对话框	
选择“Summaries for groups of cases”	选择数据表达类型
【Slices Represent】：选择“Sum of variable”	选择自定义饼图各块的含义
【Variable】：营业额	定义饼图各块代表营业额大小
【Defined Slices by】：商品类别	选择变量“商品类别”作为分块变量
单击【OK】按钮	定义完成

执行以上操作之后，生成如图 4-24 所示饼图。该图形中每一块代表了一种商品销售额，该块的占比大小即代表该商品销售额占总销售额的比例。

与 SPSS 老版本比较，自 SPSS 14.0 开始饼图引入了图例。这使图形较之以往直接在图上标注各块的名称更加规范。观察图形可以发现，由于未编辑的饼图只能直观表示各类在整体中的比例，却不能表示每类具体的值有多大。所以未编辑的饼图仅用于定性了解各类的比例而无需定量了解具体值的时候。若需要了解各类的具体取值，还要对图形进行进一步的编辑。

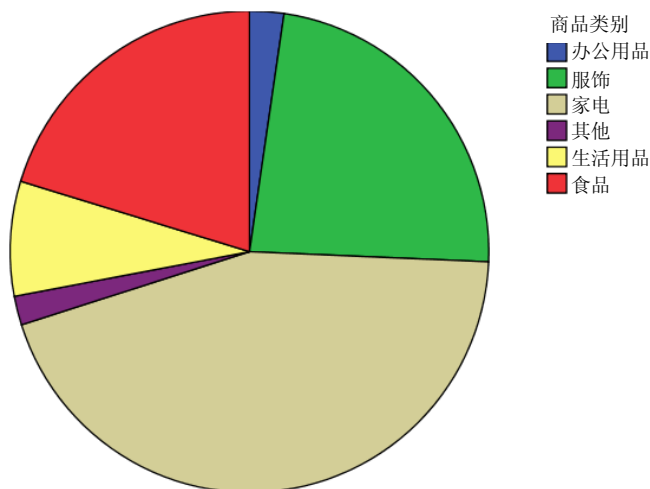


图 4-24 饼图

4.2.5 高低图（High-Low Charts）

高低图用于同时描述数据长期和短期的变化趋势。高低图是证券市场中 K 线图的主要构成部分。高低图主要分为 5 类，如表 4-5 所示。

表 4-5 高低图的主要类型

名 称	图 形	说 明
简单高低图 (simple high-low-close)		主要用于单个证券，通常以日期作为横坐标。每条线上 3 点分别代表证券价格的最高值、最低值和日收盘值
分组高低图 (clustered high-low-close)		与简单高低图类似，但是它可以同时描述两种或两种以上证券的价格情况
简单极差图 (simple range bar)		主要用于单个证券，用长条的长度代表每个时间段最高值与最低值之差
分组极差图 (clustered range bar)		与简单极差图类似，但是可以描述两个或两个以上证券的情况
对比面积图 (difference area)		描述两个现象在同一时间内相互变化的对比关系

例 4.7 绘制某股票 10 日价格高低图。利用数据文件“SFZ.sav”绘制高低图。执行以下操作：

执行【Graphs】/【Legacy Dialogs】/【high-low】命令，弹出【High-Low Charts】对话框	
选择“simple high-low-close”、 “Summaries of separate variables” 【High】：日最高价	选择绘制简单高低图，弹出如图 4-25 所示对话框 定义高低图各项

【Low】: 日最低价

【Close】: 日收盘价

【Category Axis】: 日期

选择日期作为分类变量

单击【OK】按钮

定义完成

执行以上操作之后,生成如图 4-26 所示高低图。其中每一条高低线上的最高线取值代表该股票的当日最高价,最低线取值代表该股票的当日最低价,圆圈所在值代表该股票的当日收盘价。

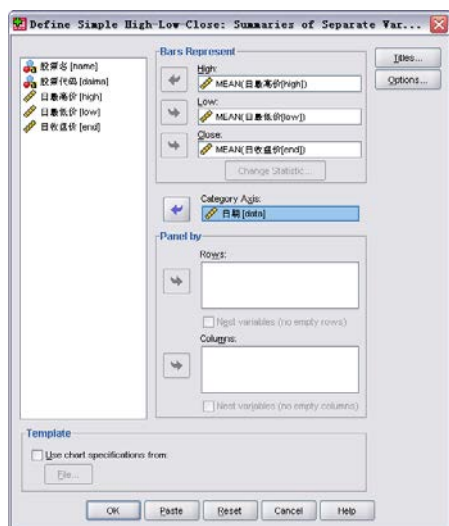


图 4-25 简单高低图定义框

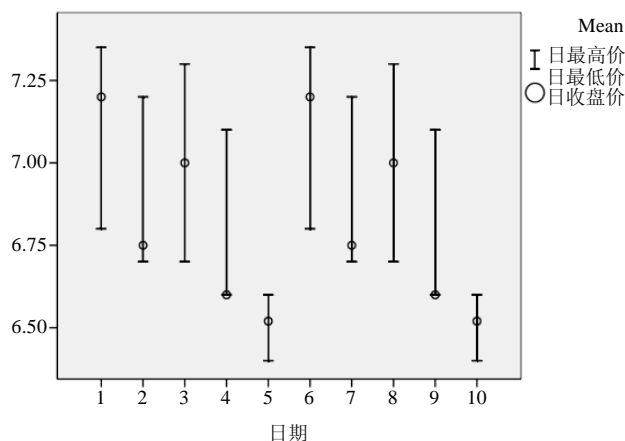


图 4-26 简单高低图

注意 在定义高低图的时候相同指标下“high”项所定义的变量值必须比“low”项所定义的变量值大,否则 SPSS 会提示出错。

4.2.6 帕累托图 (Pareto Charts)

帕累托图能把各个变量按照某一指标值从大到小降序排列,并用长条长度表示各变量该指标值的大小,同时绘制累计百分比曲线,使用户可以通过图形直观地找出各个变量的重要程度。帕累托图分为简单帕累托图和分段帕累托图两种。前者采用简单长条,后者采用分段长条,其他地方是类似的。

例 4.8 某店营业额。采用例 4.6 绘制饼图的数据文件“chaoshi.sav”,绘制该店营业额的帕累托图。

执行以下操作:

执行【Analyze】/【Quality Control】/【Pareto Charts】命令,弹出【Pareto Charts】对话框

选择“simple”、

选择绘制简单帕累托图,弹出如图

“ Counts or sums for groups of cases ”

4-27 所示对话框

【Bars Represent】:

Sums of variable: 营业额

定义帕累托图长条

【Category Axis】: 商品类别

选择商品类别作为分类变量

单击【OK】按钮

定义完成

生成如图 4-28 所示帕累托图。由图 4-28 可以看出，帕累托图是一个双轴图形。图形中某种商品对应的长条长短代表该商品的具体营业额大小，对应的曲线取值代表该商品及该商品左边所有商品的累计营业额占总营业额的比例。

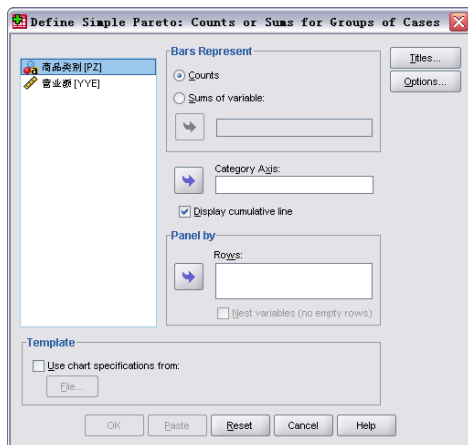


图 4-27 帕累托图定义对话框

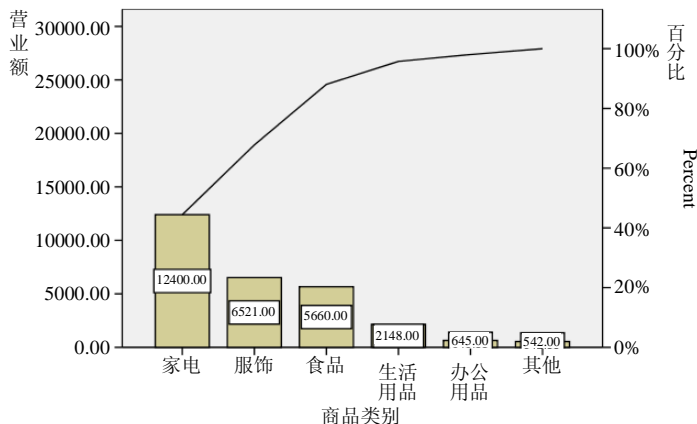


图 4-28 帕累托图

与饼图相比，帕累托图不仅可以直观反映各变量的重要程度，还给出了其具体的取值和所占比例的具体值。对于那些既需要直观描述又需要定量表达的问题通常采用帕累托图显示。既然帕累托图包含多于饼图的信息，那么为什么在实际问题中没有完全用帕累托图来替代饼图呢？这是由于饼图更加直观更加被大众所熟悉。所以，帕累托图虽然功能更强大，但是一般只用于比较专业的地方。对于一般的定性描述，采用饼图就足够了。

注意 老版本的 SPSS 将帕累托图 (Pareto Charts)、质量控制图 (Control Charts) 放在【Graphs】菜单下。SPSS 17.0 将这两个图形放在了【Analyze】菜单下的【Quality Control】子菜单中。但是在本书中, 我们仍然将这两个图形放到本章介绍。

4.2.7 质量控制图 (Control Charts)

质量控制图主要用于监测生产过程中的变化趋势, 从而提示生产者发现问题, 并采取措施来及时纠正某些不良趋势。

执行【Analyze】/【Quality Control】/【Control Charts】命令, 弹出如图 4-29 所示对话框。SPSS 中的质量控制图共分为 4 类。

- **X-Bar,R,s:** 主要包括均数—极差控制图 (X-R) 和均数—标准差 (X-s) 控制图两类。当控制图的每个小类的数据样本少于 10 个时默认为前者, 否则默认为后者。
- **Individuals,Moving Range:** 个值—移动极差控制图。如果控制图每个小类的数据样本只有一个, 则采用这种图形反映数据波动情况。
- **P,np:** 不合格品率和不合格品率控制图。
- **C,u:** 缺陷数和单位缺陷数控制图。

例 4.9 零件质量控制。利用数据文件“ljzl.sav”绘制零件控制的均数控制图和极差控制图。

执行如下操作:

执行【Analyze】/【Quality Control】/【Control Charts】命令, 弹出【Control Charts】对话框	
选择“X-Bar,R,s”、“cases are units”	选择绘制均数 - 极差图, 弹出如图 4-30 所示对话框
【Process Measurement】: weight	定义零件质量为监测变量
【Subgroups Defined by】: number	选择零件号别作为分类变量
单击【OK】按钮	定义完成

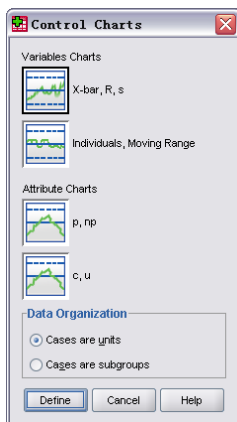


图 4-29 【Control Charts】对话框

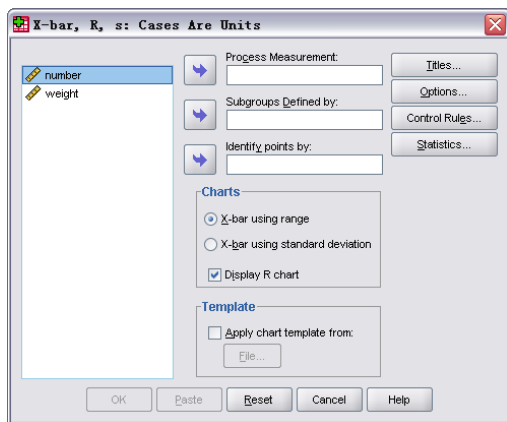


图 4-30 【均值—极差图定义】对话框

生成如图 4-31 所示均值控制图和图 4-32 所示极差控制图。

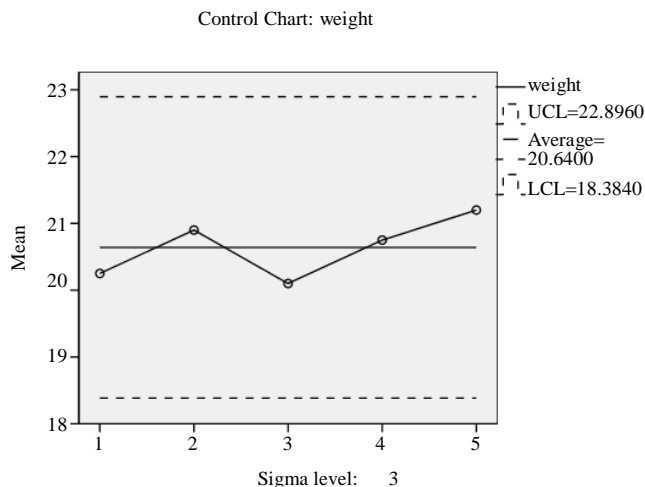


图 4-31 均值控制图

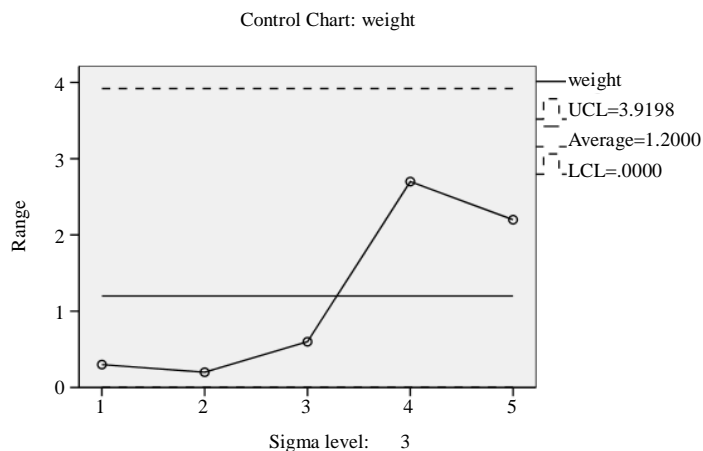


图 4-32 极差控制图

均值控制图和极差控制图反映了均值或极差的平均值及上下界限。生产过程只要保证均值和极差在控制图给定的上下界限之内波动就是正常的。如果波动范围超过了图形中所给出的界限，则应引起重视。

4.2.8 箱图（Boxplot）与误差条图（Error Bar）

箱图和误差条图都用于描述数据的分布信息。箱图主要描述数据的中位数、四分位数及极值。误差条图主要描述均值、标准差、置信区间等。二者具体的绘制过程都与条图类似，这里也不再重复介绍。

由数据文件“chengji_1”按照班级分类绘制学生语文成绩的简单箱图和误差条图分别如图 4-33 和图 4-34 所示。

在图 4-33 中，每一个箱体上方那条线的取值代表该班学生语文成绩最高分，下方那条线的取值代表该班学生语文成绩最低分。箱体自身的三条线从上到下分别代表成绩的 3/4 分位点、中位点、1/4 分位点的取值。观察图形可以发现，班级 5 比较特殊，因为它的成绩最高分就是其 3/4 分位点

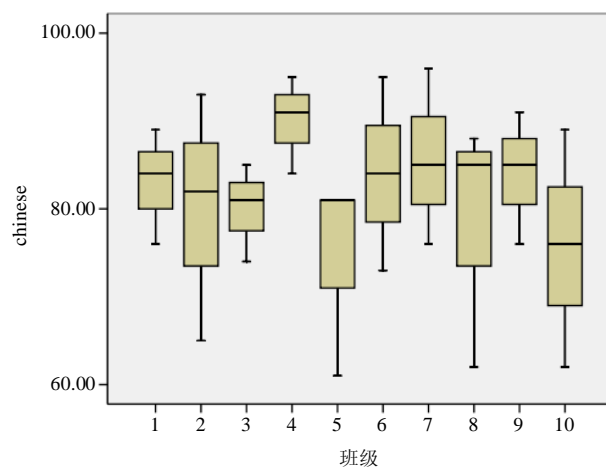


图 4-33 简单箱图

图 4-34 是选择按照置信区间画出的误差图。其中每根线条的长度代表了数据在 95% 的置信区间范围。需要注意的是误差图的定义框，除了置信区间，我们还可以选择按照标准差、标准误差来绘制误差图。

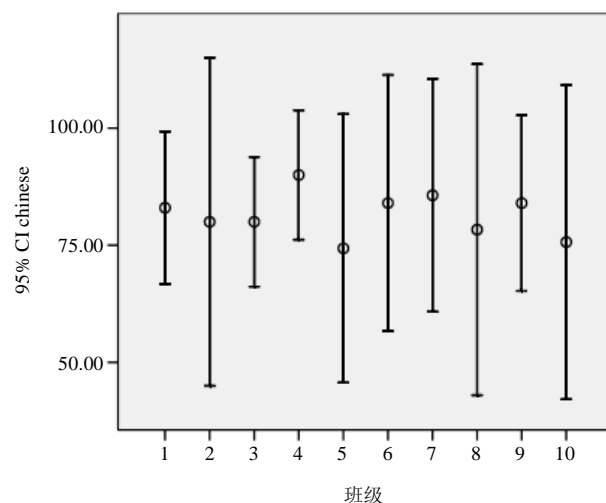


图 4-34 95% 置信区间的简单误差图

4.2.9 金字塔图 (Population Pyramid)

金字塔图是 SPSS 新增的一种图形。在社会经济学中，常常出现这样一种现象，即低收入者占人口较大比例，高收入者占人口较小比例。这就是通常所说的金字塔形。金字塔

图是一种能够很好反映这种现象分布的图形。

例 4.10 人口收入。已知有数据文件“shouru.sav”，绘制收入的金字塔图。

执行以下操作：

执行【Graphs】/【Legacy Dialogs】/【Population Pyramid】命令，弹出如图 4-35 所示对话框

【Counts】：compute counts from data	
【show distribution over】：salary	定义图形显示的是收入的分布
【split by】：sex	定义图形按照性别来划分左右
单击【OK】按钮	定义完成

生成如图 4-36 所示的金字塔图。

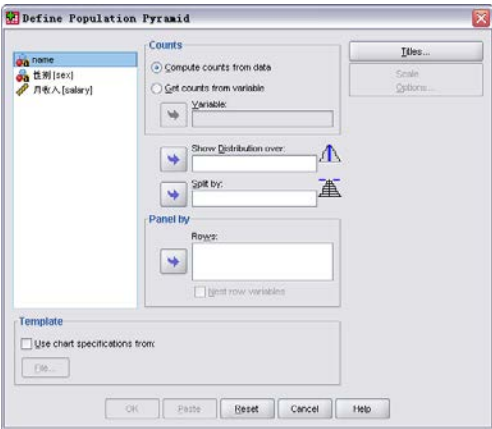


图 4-35 【金字塔图定义】对话框

图 4-36 反映了男女收入在各层次的分布。显然，低收入者数量最多，高收入者最少。需要注意的是，虽然在 SPSS 中这种图形取名为金字塔形，但这仅仅是由于它是从金字塔形的社会问题中引申出来的。在其他问题中，也完全有可能绘制出来倒金字塔、瓶形等各种各样的形状。希望读者不要局限于单一的金字塔形内。

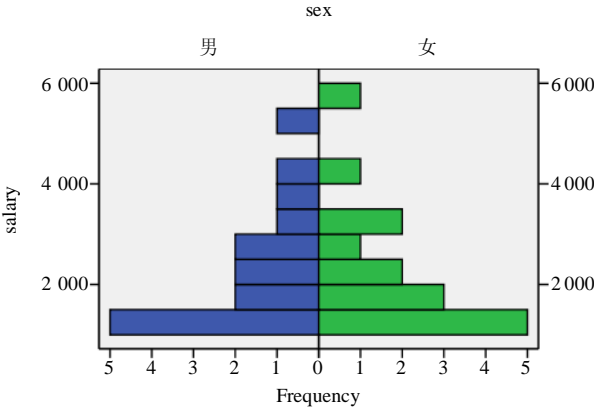


图 4-36 金字塔图

4.2.10 散点图（Scatter/Dot）

散点图是用来表示两个或两个以上变量之间相互关系的图形。在做统计分析时，要选择恰当的统计方法，通常都离不开散点图。

执行【Graphs】/【Legacy Dialogs】/【Scatter】命令，弹出散点图的定义框。如表 4-6 所示，SPSS 提供了 5 类常用散点图。

表 4-6 散点图的主要类型

名 称	图形	说 明
简单散点图 (Simple Scatter)		描述两个变量之间的相互关系
矩阵散点图 (Matrix Scatter)		利用类似矩阵的形式，在一张图上同时描述多个变量之间的两两关系
重叠散点图 (Overlay Scatter)		利用将两幅简单散点图叠加到一张图上的形式同时描述多个变量之间的两两关系
3D 散点图 (3-D Scatter)		描述三个变量之间的相互关系
简单点图 (Simple Dot)		描述一个变量在各个值的分布情况

由于各类散点图的定义界面类似，所以这里仅介绍矩阵散点图。

例 4.11 房屋销售。利用数据文件“home sales.sav”绘制变量房屋土地成本、房屋增值成本、房屋售价之间的矩阵散点图。

执行以下操作：

执行【Graphs】/【Legacy Dialogs】/【Scatter】命令，选择 Matrix Scatter 定义绘制矩阵散点图，弹出如图 4-37 所示对话框

【Matrix variables】：将要绘制的 3 个变量全部选入

选择例中要求的变量

单击【OK】按钮

定义完成

生成如图 4-38 所示的矩阵散点图。该散点图可以看成是一个 3×3 的图形矩阵。矩阵的每个元素为两个变量间的一个二维散点图。因为变量不能与自身作散点图，所以矩阵散点图的对角线元素为空。

由图 4-37 可知，散点图的定义框比前面介绍的图形多了两个特色定义框。其中【Set marks by】用来定义标记变量，即标记变量取值不同时，绘制的散点图中散点的颜色不同。【Label cases by】主要用于定义散点的标签名变量。

4.2.11 直方图（Histogram）

直方图主要用于描述变量的分布情况。它是 SPSS 中一种很常用的图形，但定义却十分简单。如图 4-39 所示，直方图定义窗口只需要在【variables】框内选择需要绘制直方图的变量，以及通过“Display normal curve”选项判断是否需要在图上画出相应的正态曲线。

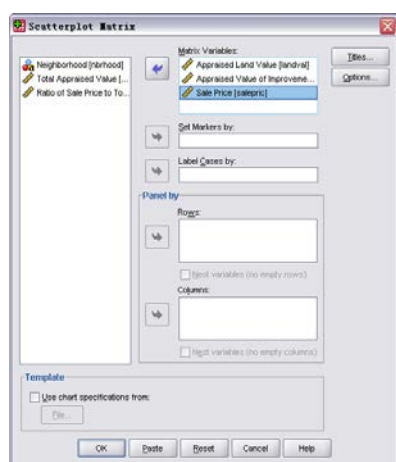


图 4-37 矩形散点图定义对话框

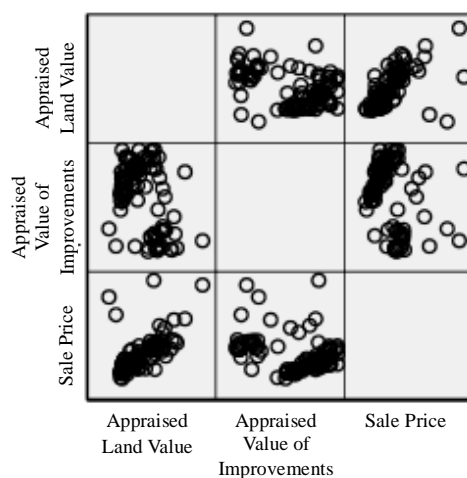


图 4-38 矩形散点图

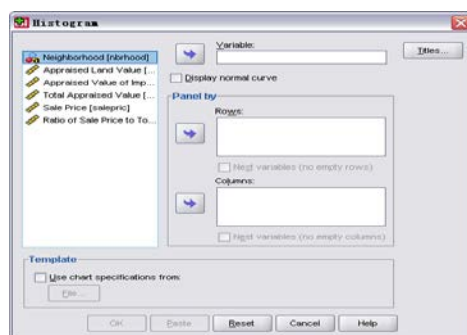


图 4-39 【Histogram】对话框

图 4-40 所示的是一个直方图的例子。如果数据样本足够大，优秀的统计工作者从图形上就可以判断变量满足的大致分布类型。

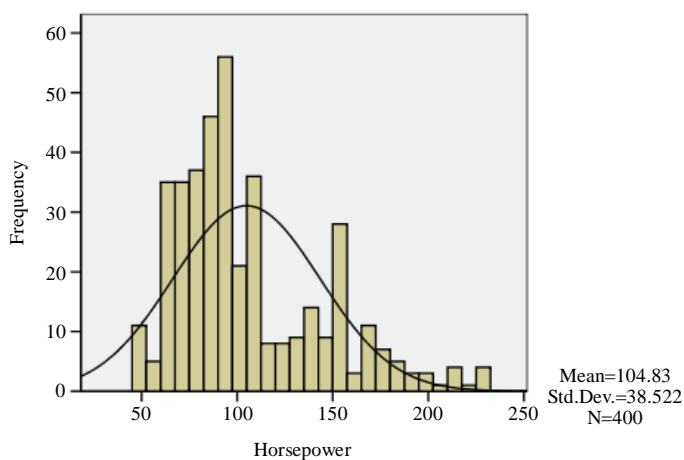


图 4-40 直方图

4.2.12 P-P图和Q-Q图

统计工作者可以用直方图来大致判断数据满足的分布类型，但是这种判断完全依赖于统计工作者的实际经验，人为主观因素太大，难免有所偏差，同时也无法判断实际分布与估计分布的差距有多大。遇到这些问题的时候，采用 P-P 图和 Q-Q 图就更客观了。

注意 老版本 SPSS 的 P-P 图和 Q-Q 图是放在【Graphs】菜单中的。在 SPSS 17.0 中，这两个图形放在【Analyze】菜单下的【Descriptive Statistics】子菜单中。

P-P 图和 Q-Q 图都是用来检验数据是否服从某种分布的辅助图形。

执行【Analyze】/【Descriptive Statistics】/【P-P Plots】命令，弹出如图 4-41 所示的 P-P Plots 对话框。

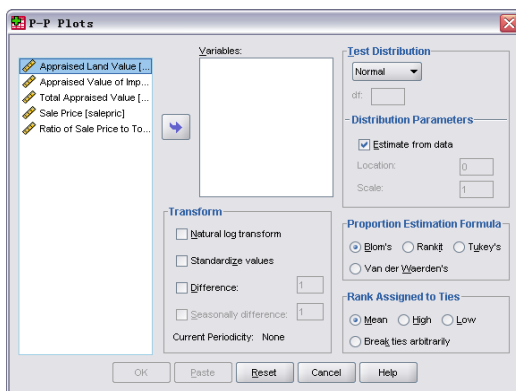


图 4-41 【P-P Plots】对话框

P-P 图的定义界面与前面的图形差别比较大，下面一一介绍。

- **【Variables】**：选择绘制 P-P 图的变量，可以同时选择多个变量。
- **【Test Distribution】**：选择待检测的分布类型。SPSS 提供了 Beta 分布、卡方分布、指数分布、伽玛分布、半正态分布、Logistic 分布、对数正态分布、正态分布等 13 种常见分布类型。下边的 df 框是用来确定 t 分布的自由度的。如果选择其他分布，该项默认不存在。
- **【Distribution Parameters】**：确定分布参数。该参数既可由 SPSS 自动从数据中估计，也可由用户自己设定。若选择“Estimate from data”，则在 SPSS View 窗口中通过表格输出估计参数值。
- **【Transform】**：定义数据的转换方式。包括取自然对数、标准化（Z 变换）、取差分（取原值与 n 个相邻值的差值替代原值）和季节差分（取原值与 n 个时期值的差值替代原值）4 种形式。当然，对于不需要转换的数据不选此项。
- **【Proportion Estimation Formula】**：定义计算预期正态概率值的方法。
- **【Rank Assigned to Ties】**：指定对相同的多个变量值的处理方式。

下面利用一组由 SPSS 自动生成的服从 $N(0, 1)$ 分布的数据绘制其关于正态分布的 P-P 图，结果如图 4-42 和图 4-43 所示。

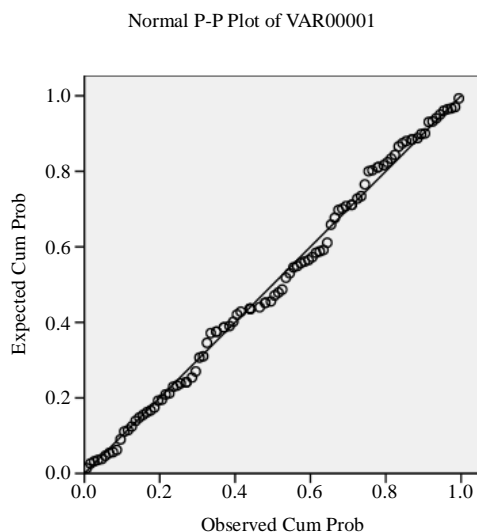


图 4-42 P-P 图

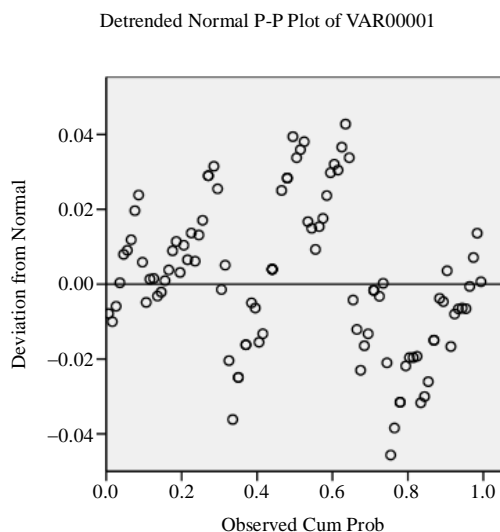


图 4-43 P-P 去势图

由于绘制图形时所选取的数据本来就服从 $N(0, 1)$ ，所以图形效果是十分好的。在 P-P 图中，检验数据是否较好地服从给定分布的标准有两个。第一，看 P-P 图上的数据点与直线的重合度。第二，看 P-P 去势图上的点是否关于直线 $Y=0$ 在较小的范围内上下波动。

Q-Q 图与 P-P 图的定义类似。二者的区别是 P-P 图比较的是真实数据和待检验分布的累计概率。而 Q-Q 图比较的是真实数据与待检验分布的分位点值。

图 4-44 和图 4-45 是利用同一组由 SPSS 自动生成的服从 $N(0, 1)$ 分布的数据绘制的其关于正态分布的 Q-Q 图与 Q-Q 去势图。从图形形态上看，它们正好和 P-P 图及 P-P 去势图对应。Q-Q 图检验数据是否较好地服从给定分布的标准也是同 P-P 图类似的两点，所以此处不再复述。

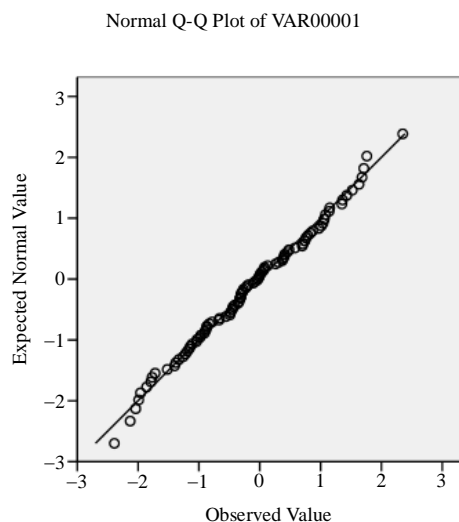


图 4-44 Q-Q 图

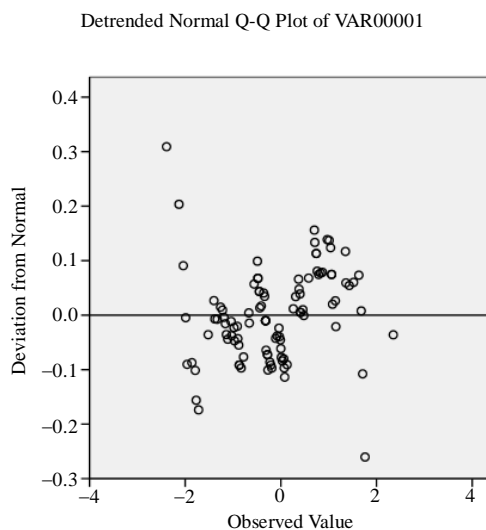


图 4-45 Q-Q 去势图

4.2.13 ROC曲线

ROC 曲线是二元判决中用来比较判决方法优劣的一种曲线。它以 p_f 做横轴， p_d 做纵轴所生成。其中 p_f 表示本来是假被误判为真的概率即虚警概率， p_d 表示本来是真判别其为真的概率，即漏检概率的补充。在实际问题中，我们说一种判别方法优于另一种方法，从 ROC 曲线上讲就是使其 p_f 尽可能小而 p_d 尽可能大，即较好方法的 ROC 曲线应该始终位于较差方法的 ROC 曲线的左上方。

注意 老版本 SPSS 的 ROC 曲线放在【Graphs】菜单中。在 SPSS 17.0 中，ROC 曲线放在【Analyze】菜单中。

例 4.12 仪器观测准确度的比较。已知数据文件“sensor.sav”，其中变量“fact”表示真实情况，变量“sensor1”、“sensor2”表示传感器关于真实数据在某一指标下的观测值。通过绘制 ROC 曲线比较两个传感器的优劣。

执行以下操作：

执行【Analyze】/【ROC Curve】命令，弹出如图 4-46 所示对话框

【Test Variables】: sensor1、sensor2

选择“sensor”作为检测变量

【State Variables】: fact

选择“fact”作为状态变量

【Value State Variables】: 1

定义“fact”取值为 1 时结果为真

【Display】: 选中前两项

绘制 ROC 曲线并显示对角线作为标识曲线

单击【OK】按钮

定义完成

生成如图 4-47 所示 ROC 曲线。对算法有兴趣的读者，可以研究 SPSS 的帮助文档。如果单纯地从 ROC 曲线的图形上看，传感器 2 的效果优于传感器 1。

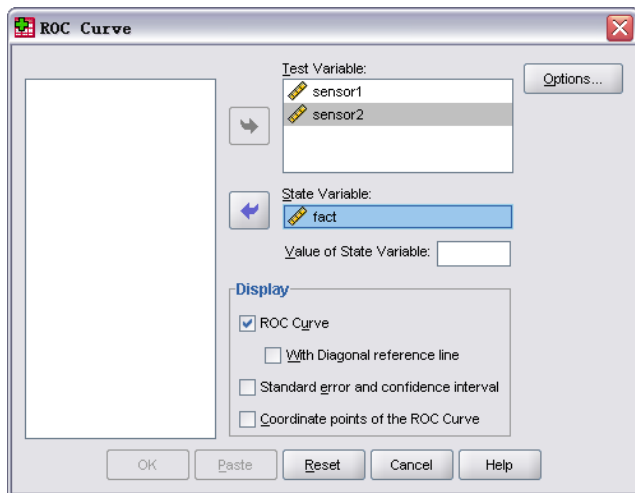


图 4-46 【ROC Curve】对话框

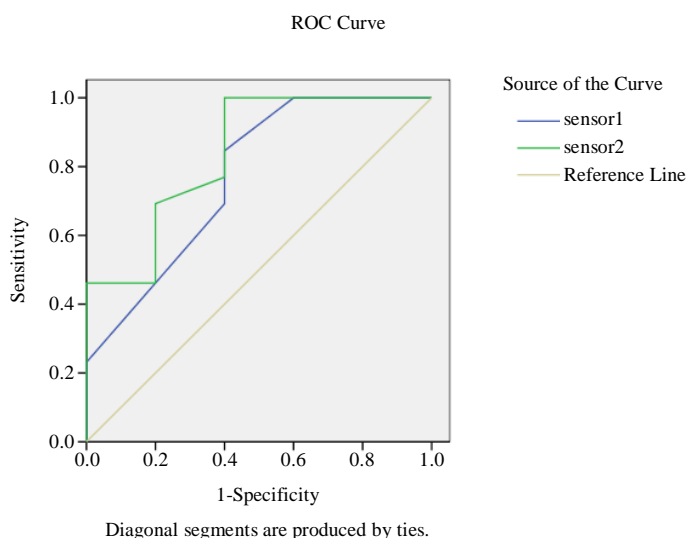


图 4-47 ROC 曲线

在图 4-46 所示的对话框中，除了例 4.12 中用到的检测变量定义框【Test Variables】、状态变量定义框【State Variables】、状态真值定义框【Value State Variables】、【Display】的前两项之外，还有如下三项。

- “Standard error and confidence interval”

计算 ROC 曲线下面积的标准误差和置信区间。

- “Coordinate points of the ROC Curve”

通过表格输出 ROC 曲线上各点的坐标值。

- 【Options】

定义 ROC 曲线计算的方法。单击图 4-46 上的【Options】按钮，弹出如图 4-48 所示的对话框，各项含义如下。

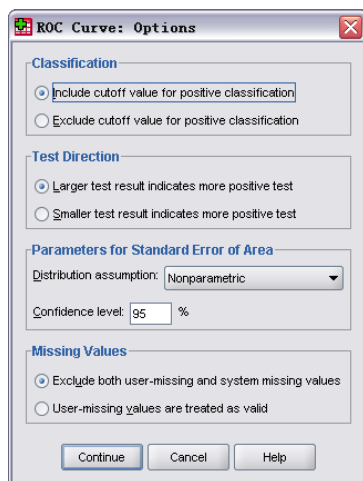


图 4-48 【Options】对话框

- ① Classification: 定义边界值的分类。
- ② Test Direction: 定义检验方向。包括检验结果越大越趋近于真和结果越小越趋近于真两种情况。
- ③ Parameters for Standard Error of Area: 定义估计曲线下面积的标准误差的方法和选择置信度。
- ④ Missing Values: 定义缺失值的处理方式。

注意 ROC 曲线在检测中有着广泛的应用,但是不能只是单纯地依赖于 ROC 曲线来比较两种方法,最后还必须通过假设检验才能得出结论。这也是初学者容易忽略的地方。

4.2.14 时间序列图 (Time Series Charts)

顾名思义,时间序列图是研究与时间序列相关的数据特征的图形。在 SPSS 17.0 中,时间序列图放在【Analyze】菜单下的【Forecast】子菜单中,共有如下三类。

- Sequence: 普通序列图。主要用于描述一个或几个变量随着另一个变量变化的趋势。
- Autocorrelations: 自相关时间序列图。主要用于研究同一变量的前一时间周期与后一时间周期对应观测点之间的相互关系。
- Cross-correlations: 互相关时间序列图。主要用于研究多个序列在对应观测点之间的相互关系。

下面我们分别举例介绍普通序列图 and 自相关时间序列图。

1. 普通序列图 (Sequence Charts)

普通序列图主要用于描述一个或几个变量随着另一个变量变化的趋势。执行【Analyze】/【Forecast】/【Sequence Charts】命令,弹出如图 4-49 所示的普通序列图定义对话框。

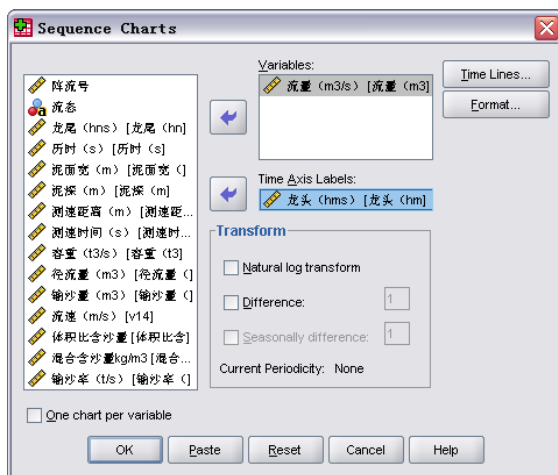


图 4-49 【Sequence Charts】对话框

普通序列图定义对话框的各项功能介绍如下。

- 【variables】

定义因变量，默认为图形纵轴，可以同时选择多个变量。

- 【Time Axis Labels】

定义时间变量，默认为图形横轴。

- 【Transform】

定义数据转换方式。

- “One chart per variable”

当【variables】框选择多个变量时，选中该项则一个变量对应一张图。否则，多个变量在一张图上。

- 【Time Lines】

定义横轴上的标识线。单击【Time Lines】按钮，弹出如图 4-50 所示对话框。

图 4-50 从上到下依次代表了三种标识线的定义方式。

① No reference lines: 不定义标识线。

② Line at each change of: 当【Reference Variable】框选中变量发生变化时生成标识线。

③ Line at data: 在【Observation】框中给定的时间处生成标识线。

- 【format】

定义图形特征。单击图 4-49 所示的【format】按钮，弹出如图 4-51 所示对话框。该对话框主要由以下三部分组成。

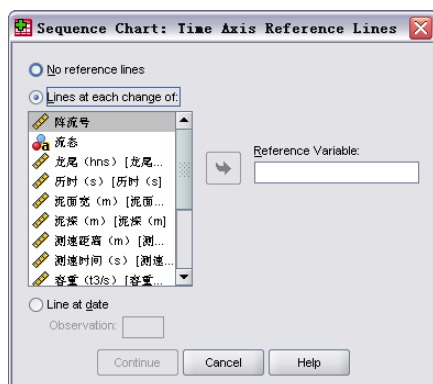


图 4-50 【标识线定义】对话框

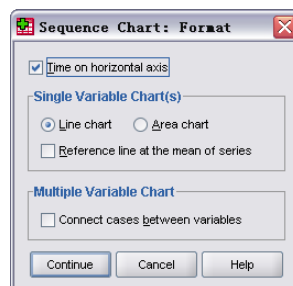


图 4-51 【图形特征定义】对话框

① Time on horizontal axis: 默认选择取时间变量为横轴，否则为纵轴。

② Single Variable Chart: 对于单变量图形，定义绘制线图还是面积图，SPSS 默认为绘制线图。同时可以选择是否显示均值的标识线。

③ Multiple Variable Chart: 选择是否显示各变量之间的差别。

下面给出一个普通序列图的例子。

例 4.13 流量随时间的变化趋势。已知数据文件“liuliang.sav”是一次突发性流体的实际观测数据，绘制流量关于龙头时间的变化趋势。

执行以下操作：

执行【Analyze】/【Forecast】/【Sequence Charts】命令，弹出如图 4-49 所示对话框

【variables】：流量

选择“流量”作为因变量即纵轴

【Time Axis Labels】：龙头

选择“龙头”作为自变量即横轴

单击【OK】按钮

定义完成，生成如图 4-52 所示图形

对于单变量图形，SPSS 默认为绘制如图 4-52 所示的普通序列线图。如果在如图 4-51 所示的【Format】对话框中选择“Area Chart”，则生成如图 4-53 所示的普通序列面积图。对于多变量图形，则只能绘制线图。

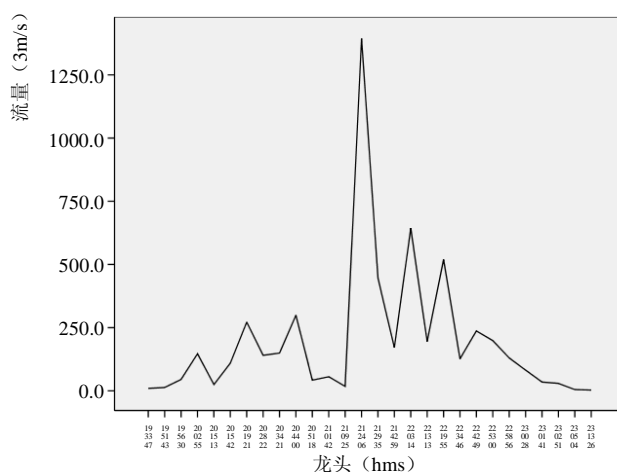


图 4-52 普通序列线图

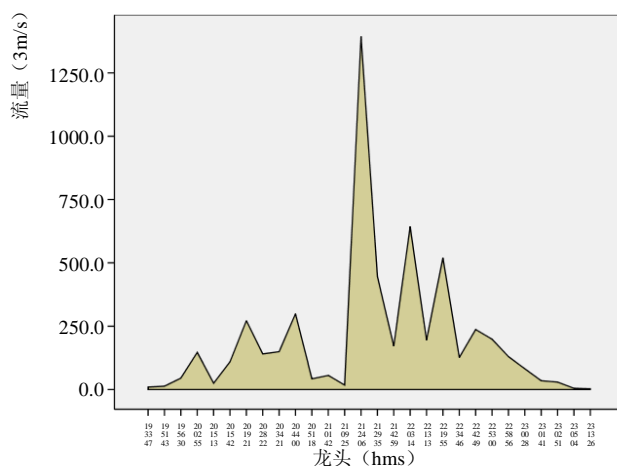


图 4-53 普通序列面积图

2. 自相关时间序列图 (Autocorrelations Chart)

自相关时间序列图主要用于研究同一变量的前一时间周期与后一时间周期对应观测点之间的相互关系。下边以一个例子来介绍自相关时间序列图。

例 4.14 某公司每年各季度销售情况。已知有某公司三年各季度的销售数据文件“time_1.sav”，试绘制该公司销售额的时间序列图，判断其是否存在周期性变动。

由于只是研究销售额，所以选择绘制自相关时间序列图，执行以下操作：

执行【Analyze】/【Forecast】/【Autocorrelations】命令	弹出如图 4-54 所示对话框
【Variables】：总销售额	选择“总销售额”作为时间序列变量
单击【Options】按钮	弹出如图 4-55 所示对话框
【Maximum Number of Lags】：4	定义以“年”为一个时间序列周期
单击【OK】按钮	定义完成

执行以上操作之后生成如图 4-56 所示时间序列图。

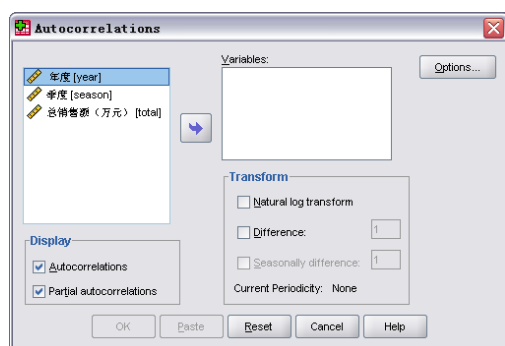


图 4-54 【Autocorrelations】对话框

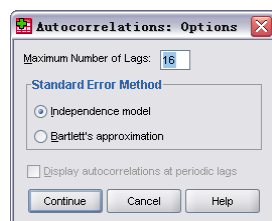


图 4-55 【Options】对话框

在时间序列图中，纵坐标代表各周期对应观测值之间的相关系数。如果相关系数为零或者为负值，表示时间序列的趋势保持原状。相关系数的最大正值若出现在最后一个时间点之前的任意时刻表示趋势变动。在图 4-56 中，前三个时间点相关系数为负或者接近于零，相关系数最大值出现在最后一个时间点。由此我们可以判断该公司的销售额是以年为周期的变动趋势。

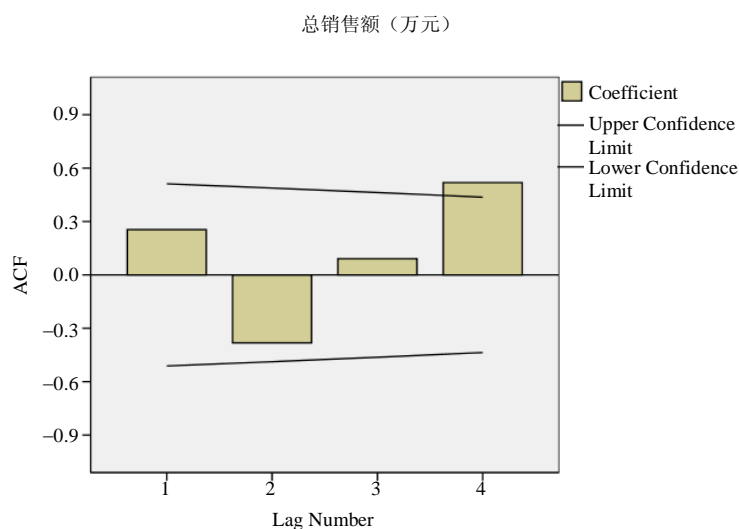


图 4-56 自相关时间序列图

3. 互相关时间序列图 (Cross-correlations Chart)

互相关时间序列图 (Cross-correlations Chart) 主要用于研究多个序列在对应观测点之间的相互关系。其定义对话框和自相关时间序列图的定义对话框几乎完全一致。这里就不再介绍了。

4.3 SPSS图形编辑

在 SPSS 中一个完整的图形创建过程包括创建数据文件、绘制统计图形和修饰所生成的图形三部分。上一节详细介绍了【Graphs】菜单下所有图形的创建。在本节中，将以条图为例介绍 SPSS 图形的编辑。

4.3.1 图形编辑概述

现在以例 4.4 生成的分段条图为例介绍图形的编辑功能。

例 4.15 编辑图 4-18。已知在【SPSS Viewer】中生成了如图 4-18 所示的分段条图，左键双击图形即进入如图 4-57 所示的图形编辑窗口。

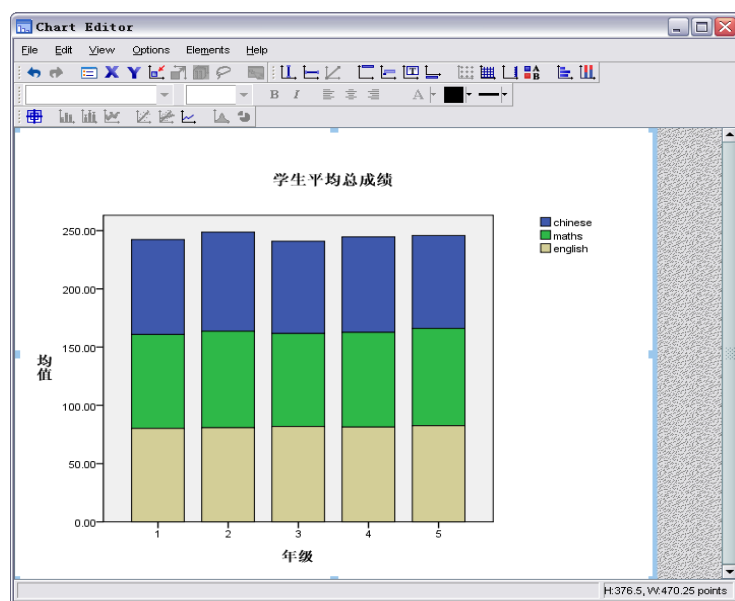


图 4-57 图形编辑窗口

现在通过执行以下操作，编辑图形：

鼠标左键单击选中条图的长条

执行【Edit】/【properties】命令，弹出如图 4-58 所示的【properties】对话框

切换到“Depth&Angle”选项卡，【Effect】/【3-D】 设置将条图转换为 3D 形式
【Angle】栏微调 调整 3D 图的条块宽度

切换到“Fill&Border”选项卡，【Pattern】

设置条图的填充格式

单击【Apply】按钮

完成图形基本设定

执行【Elements】/【Show data labels】命令

显示图形上各段取值

执行以上操作之后图形转换为如图 4-59 所示的形式。

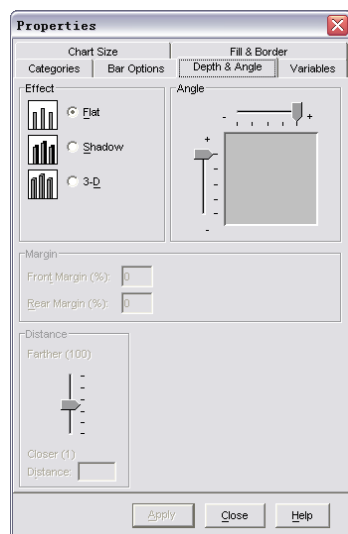


图 4-58 【properties】对话框

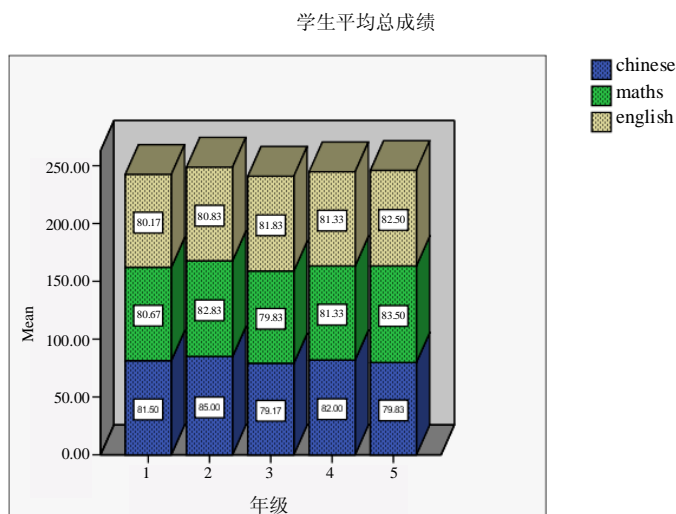


图 4-59 编辑后的条图

与 SPSS 以前的版本相比，SPSS 17.0 的图形编辑窗口的菜单变得简洁明了。虽然菜单类减少了，但是编辑功能并没有因此减少。各菜单项的主要作用如下。

- ①【File】：保存和打开定义好的图形模板。图形模板是指保存用户自定义的图形大小、颜色等各类信息的文件。
- ②【Edit】：编辑图形的大小和颜色、横纵坐标的显示方式等图形信息。
- ③【View】：定义视图窗口。
- ④【Options】：编辑图形的框架特征，即标识线、标题、脚注等信息。
- ⑤【Elements】：编辑具体的图形元素。
- ⑥【Help】：帮助菜单。

由于【File】、【View】、【Help】菜单比较简单，所以仅详细介绍【Edit】、【Options】、【Elements】三个图形编辑的主要菜单。

4.3.2 图形基本设定——Edit菜单

【Edit】菜单如图 4-60 所示，其主要功能为编辑图形基本信息和编辑 X、Y、Z 轴基本信息。

首先选择要编辑的对象，有两种选择方式。例如，要编辑 X 轴，执行【Edit】/【Select X Axis】命令，弹出如图 4-61 所示对话框。或者直接用鼠标双击 X 轴，也会进入相同的编辑对话框。

与 X 轴的编辑对话框类似，Chart 编辑对话框、Y 轴编辑对话框和 Z 轴编辑对话框也是由如图 4-61 所示的多个选项卡构成，每个选项卡有着特定的作用。由于这 4 个编辑对话框的很多选项卡是重复的，所以就直接通过表 4-7 将 4 个编辑框中各主要选项卡的功能列举出来。

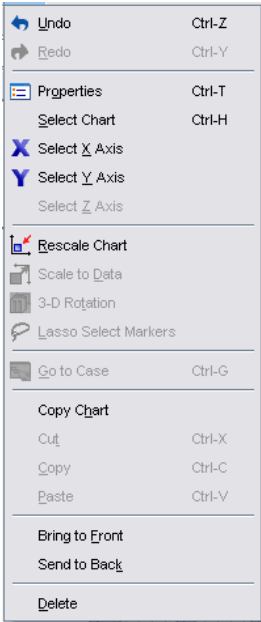


图 4-60 图形编辑窗口的 Edit 菜单

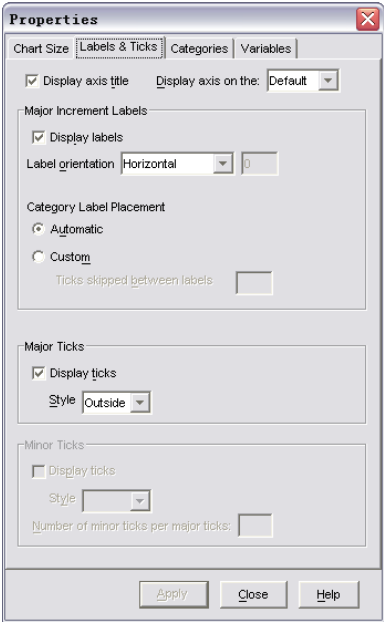


图 4-61 【X 轴的编辑】对话框

表 4-7 Edit 菜单下编辑对话框的选项卡

名 称	功 能
【Chart size】	定义图形大小
【Fill & Border】	定义图形填充色、边界线条颜色、边界线图宽度等
【Variables】	显示图形上的变量信息，并且可以通过拖曳方式调整变量在图形上的位置
【Labels & Ticks】	定义坐标、坐标刻度、坐标值的显示方式等
【Number format】	定义坐标值的数据格式
【Scale】	定义坐标轴的刻度
【Panel】	定义图组的显示方式

【Edit】菜单所定义的图形基本信息可以通过执行【File】/【Save chart template】命令将其保存为一个模板。其他图形直接套用已经定义好的模板即可。

4.3.3 图形高级设定——Options菜单和Elements菜单

【Options】菜单主要用于设定图形的框架特征，表 4-8 列出了其主要选项的功能。

表 4-8 【Options】菜单主要选项功能

名 称	功 能
【X axis references line】	定义 X 轴标识线
【Y axis references line】	定义 Y 轴标识线
【References line from equation】	从方程中获取标识线
【Title】	编辑图形标题
【Annotation】	编辑图形注释
【Text Box】	编辑图形内部文本文字格式
【Footnote】	编辑图形脚注
【Show chart in the diagonal】	把图形按照对角线形式显示
【Show grid lines】	显示/隐藏图形的网格剖分线
【Show derived lines】	在图形左、右两边同时显示纵轴坐标
【Hide legend】	显示/隐藏图例
【Transport chart】	图形转置

在图形编辑中，【Options】菜单侧重于编辑图形整体上的一些特征，而图形的一些细节的修饰就依赖于【Elements】菜单了。表 4-9 列出了【Elements】菜单的主要选项功能。

表 4-9 【Elements】菜单主要选项功能

名 称	功 能	适用图形
【Data label mode】	在图形上标识某一点的数据值	广泛适用
【Show data labels】	显示/隐藏某一点的数据值	广泛适用
【Show error bars】	显示/隐藏误差条图	高低图、帕累托图、误差条图、金字塔图
【Show line Markers】	显示/隐藏线条标记符号	线图、面积图、高低图、帕累托图、质量控制图、P-P 图、Q-Q 图、简单序列图、时间序列图
【Fit line at Total】	在全局拟合图形的回归曲线。包括直线、二次曲线、三次曲线、均值标识线、局部加权回归直线 5 种拟合方式	散点图、P-P 图、Q-Q 图
【Fit line at subgroups】	按照分组拟合曲线	散点图
【Interpolation line】	在图形中添加线条将图形中元素连接起来，包括折线、步进线、跃进线、光滑曲线 4 种可选线条	线图、高低图、帕累托图、质量控制图、误差条图、散点图、P-P 图、Q-Q 图、简单序列图、时间序列图
【Explode Slice】	将饼图某块突出显示	饼图

图形的编辑基本上是通过【Edit】、【Options】和【Elements】菜单完成的。不同的图形在编辑的过程中都有自己的特色命令。如果能够熟练运用这些编辑命令，用户一定能够做出自己喜爱的统计图形。

4.4 交互式统计图形

前面介绍了常用统计图形的绘制和编辑，本节主要通过交互条图的绘制和编辑来简单介绍交互式统计图形。

4.4.1 交互式统计图形概述

交互式统计图形是 SPSS 一项强大的绘图功能。如图 4-62 所示，它包括交互式条图、点图、线图、带状图、点一线图、面积图、饼图、箱图、误差条图、直方图、散点图这 11 种类型。

利用交互式统计图形可以绘制出各式特色图形。同一般的统计图形相比，交互式的统计图形的操作主要是利用鼠标拖曳方式，图形更加多样，图形编辑功能也更加强大。但是，由于交互式统计图形比一般的统计图形更加精致，所以其所占用的存储空间更大。同时也对计算机的硬件设备要求较高。

由于交互式统计图形的本质和常用统计图形一致，所以这里就只介绍交互式条图。

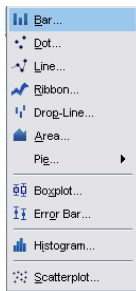


图 4-62 交互式统计图形的类型

4.4.2 交互式条图的界面

执行【Graphs】/【Legacy Dialogs】/【Interactive】/【Bar】命令，弹出如图 4-63 所示的【交互式条图】创建对话框。该对话框共有 5 个选项卡，分别介绍如下。

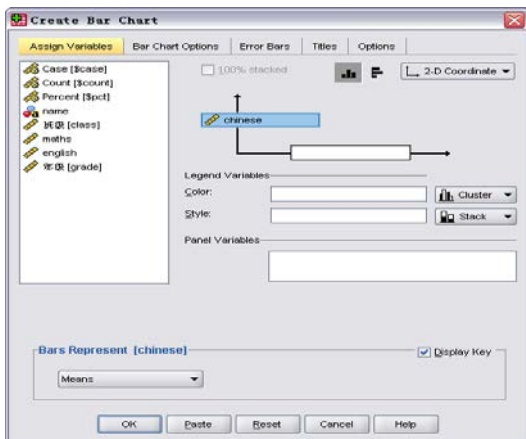


图 4-63 【交互式条图】对话框

• 【Assign Variables】

如图 4-63 所示，该选项卡用于设置图形的变量信息。主要包括图 4-64～图 4-66 所示的 3 个功能组。

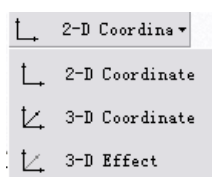


图 4-64 选择坐标



图 4-65 定义坐标变量

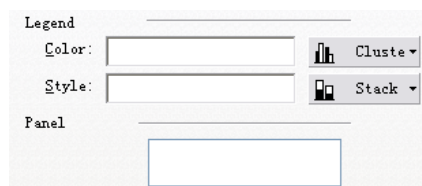


图 4-66 定义分类变量

图 4-64 用来定义坐标轴。共包括三类，分别表示二维坐标、三维坐标和 3-D 效果的二维坐标。图 4-65 表示选择坐标的变量。通过鼠标拖曳的方式将图 4-63 变量框中的变量移到对应的坐标轴位置。其中横轴代表分类变量，纵轴代表汇总变量。选择了汇总变量之后还可以通过图 4-63 下方的【Bar represent】下拉列表选择条图长条所代表的变量统计意义。图 4-66 用来定义分组变量的区分标准。包括按照颜色和风格来区分两种方法。【Panel】则用来定义绘制多张统计图形的分块变量。

注意 “Style”和“Panel”栏只能添加名义测量。如果不是名义测量，SPSS 将提示是否将该变量转换为名义测量。

• 【Bar Charts Options】

图 4-67 所示选项卡用于定义条图长条的形状、显示信息等。



图 4-67 【Bar Chart Options】对话框

- ① **Bar Baseline:** 定义长条基线值。当大于基线值时长条向上，小于基线值时长条向下，其中：

Automatic: 系统自动设定该基线值的大小；

Custom: 用户自定义该基线值的大小；

- ② **Bar Labels:** 定义长条标签显示信息，包括显示变量数和变量值两种情况。

• 【Error Bars】

误差条图选项卡。定义是否显示误差条图以及误差条图的置信区间、形状、方向等。

• 【Titles】

定义标题、脚注等。

- 【Options】

定义图形的外围特征。

4.4.3 交互式条图实例

例 4.16 交互式条图。利用数据文件“Employee Data.sav”绘制交互式条图，并比较它和普通统计图形的区别。

执行以下操作：

执行【Graphs】/【Legacy Dialogs】/【Interactive】/【Bar】命令，弹出如图 4-63 所示对话框

单击【Assign Variables】选项卡

选择 3-D Effect 图形，其中 X 轴选入 gender，Y 轴选入 salary

【Color】：educ

选择按照颜色区分雇员的受教育程度

【Panel】：minority

按照是否是少数民族绘制两张统计图形

单击【OK】按钮

定义完成

生成如图 4-68 所示图形。

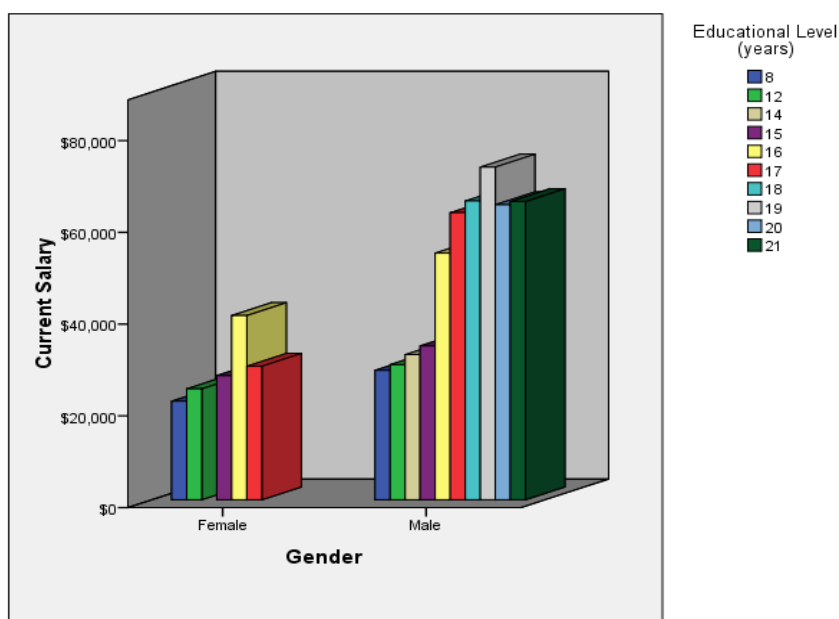


图 4-68 交互式条图

从图 4-68 可以明显看出，男性雇员比女性雇员薪酬水平高。同时由于选择了“educ”作为色彩标签变量，因此很容易利用长条的颜色来比较不同受教育程度人员的薪酬水平。从图形上看出，无论是男性还是女性，薪酬水平最高的都不是受教育年限时间最长的，反而是受教育年限次高的那些人员。

对于交互式条图，仍然可以对其进行编辑。双击图形，弹出如图 4-57 所示的图形编辑界面。对图形进行编辑之后，生成了如图 4-69 所示新图形。对图 4-68 做图形转置、图形背景和色彩优化、添加男女薪水最高值等编辑，编辑后的交互式条图如图 14-69 所示。

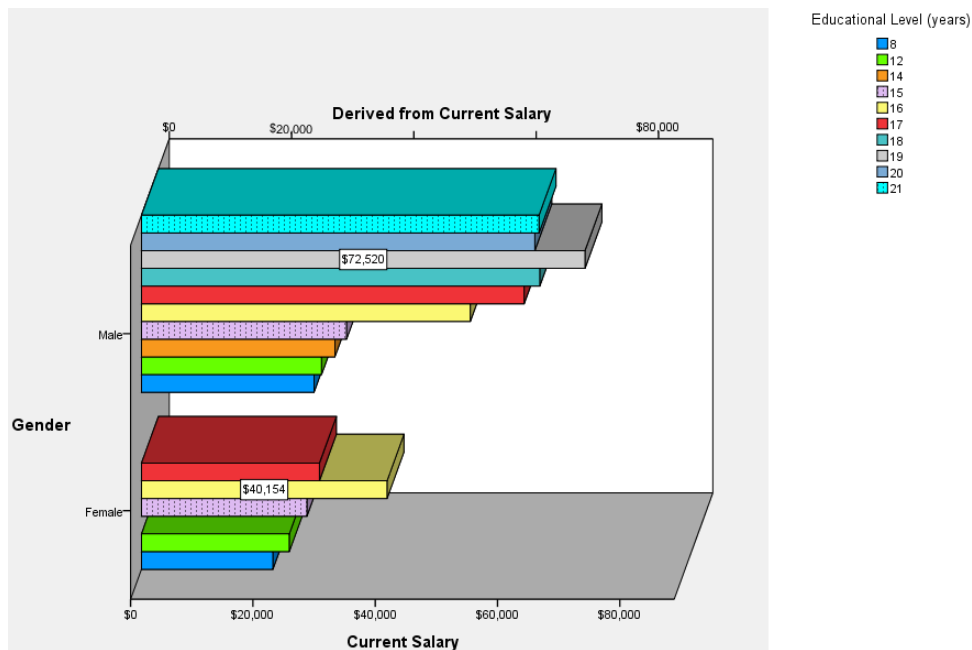


图 4-69 编辑后交互式条图

交互式图形可以绘制出非常精美细致的图形，且具有十分强大的功能。不论是其他的统计软件还是 SPSS 中普通统计图都很难达到这种既美观且实用的效果。不过需要注意的是，此时生成的交互式统计图对打印机的分辨率要求也比较高。一般的打印机可能无法打印出这么细致的图形。

4.5 本章小结

本章介绍了 SPSS 统计图形绘制。统计图形的绘制共包括 4 种方法：

- 利用图形生成器绘制图形；
- 利用图形模板选择器绘制图形；
- 绘制常见统计图；
- 绘制交互式统计图。

其中，需要掌握常见统计图的作用和绘制方法。同时，学习利用图形生成器和模板选择器来简化绘图过程，并了解通过图形编辑或者交互式绘图来优化图形。

第 5 章 SPSS 报表

上一章全面介绍了 SPSS 中的统计图形，本章介绍 SPSS 的报表。SPSS 的所有统计结果都是以表格的形式输出到结果浏览窗口的，但这里介绍的报表是比统计分析结果更简单明确的数据列表。SPSS 的报表功能是以表格的形式，按照一定的要求对数据进行列表以表现数据内在的联系。当然，这里的介绍是建立在读者已知数据文件的建立、编辑和整理方法的基础上的。本章的主要内容包括：

- 简单记录报表——Reports 子菜单
- 高级报表——Tables 子菜单

5.1 简单记录报表——Reports 子菜单

在【Analyze】菜单下的子菜单【Reports】是与简单记录报表有关的功能，它还可以计算一些简单的描述统计量。它的每个过程如图 5-1 所示。

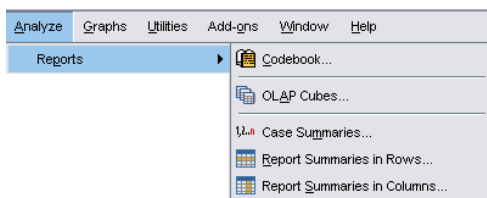


图 5-1 【Reports】子菜单

如图 5-1 所示，【Reports】子菜单共由 5 个过程构成。其中【Codebook】过程主要用来输出选择变量的基本信息，包括变量的名字、标签、类型、格式、测量尺度以及变量取值的均值、方差、分位数等等。由于其操作非常简单，本节就不做单独介绍了。

本节将详细介绍【Reports】子菜单的其余 4 个过程。本章例子如果没有特殊说明，仍然是以光盘中的数据文件“Employee Data.sav”为例。

5.1.1 在线分析处理——OLAP 过程

SPSS 的【OLAP】过程即为在线分析处理过程，它可以按照一个或几个变量的每个分组形成分层的报表，在报表中对所选的另一些变量进行相关的统计分析，这个过程也称为分层报告过程。

执行【Analyze】/【Reports】/【OLAP Cubes】命令，弹出如图 5-2 所示的【OLAP Cubes】对话框，下面介绍其中的元素。

1. 【Summary Variable】框

汇总变量框，其中放置的变量是将进行汇总分析的变量。

注意 其中的变量必须是数值型变量，最好是 Scale 测度水平定义的连续变量。

2. 【Grouping Variable】框

分组变量框，其中的变量是分组变量，将按其中的变量对全部观测量进行分组。变量可以是数值型也可以是字符型。

3. 【Statistics】子对话框

【统计量】子对话框。单击【Statistics】按钮，弹出如图 5-3 所示的子对话框，主要用于定义分层报告表中的统计量。在这个对话框左侧的“Statistics”框中列出了 SPSS 所有的统计量，它们的意义列在表 5-1 中；右侧的“Cell Statistics”框列出了所有将在报表中使用的统计量，图 5-3 所示是系统的默认状态，可以用中间的箭头修改其中的设置。

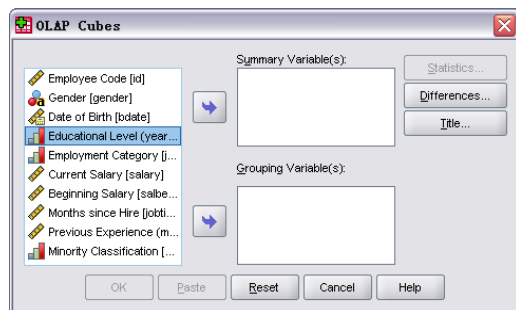


图 5-2 【OLAP Cubes】对话框

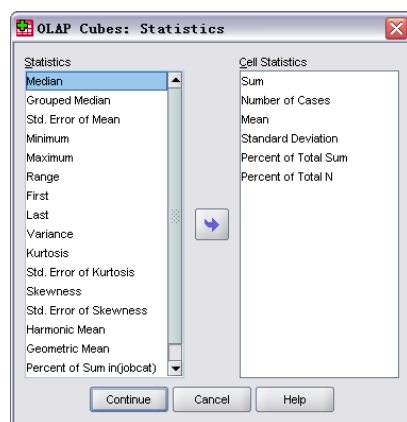


图 5-3 【Statistics】子对话框

4. 【Differences】子对话框

【差值】子对话框。单击【Differences】按钮，弹出如图 5-4 所示的子对话框，它主要用于对报表中差值计算的设定。

• “Differences for Summary Statistics” 单选框

该单选框有 3 个选项：

None，系统默认选项，意思是不计算差值；

Differences between variables，计算汇总变量之间的差值，如果选择了这一项，则需要再在“Differences between Variables”框内进一步设定参数；

Differences between groups，计算分组变量内两组之间的差值，如果选择了这一项，则需要再在“Differences between Groups of Cases”框内进一步设定参数。

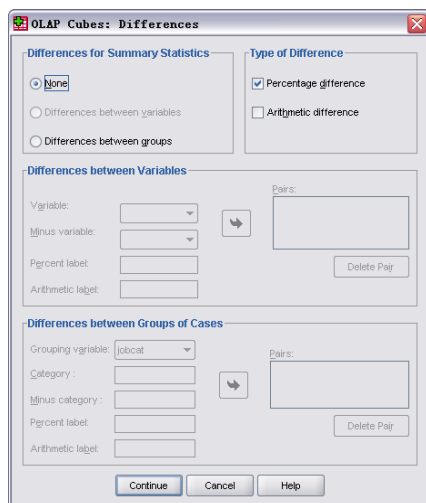


图 5-4 【Differences】子对话框

- “Type of Differences”复选框

包含两个选项：

Percentage difference，百分比差值，即输出所配对的变量或组对的差值与被减的变量或组对的百分比；

Arithmetic difference，算术差值，即输出所配对变量或组对的差值。

表 5-1 SPSS 统计量意义

统 计 量	统计量意义	统 计 量	统计量意义
Sum	和	Kurtosis	峰态
Mean	均值	Skewness	偏态
Standard Deviations	标准差	Geometric Mean	几何平均值
Number of Cases	观测量数目	Harmonic Mean	调和平均值
Median	中位数	Std. Error of the Mean	平均标准误差
Grouped Median	组中位数	Std. Error of Kurtosis	峰态标准误差
Maximum	最大值	Std. Error of Skewness	偏态标准误差
Minimum	最小值	Percentage of Total N	观测量总数百分比
First	第一个观测量值	Percentage of Total Sum	观测量总和百分比
Last	最后一个观测量值	Percentage of N in	组变量观测量总数所占百分比
Range	极差	Percentage of Sum in	组变量观测量值占总和百分比
Variance	方差		

- “Differences between Variables”框

变量差值框，用于设定变量差值计算的参数。在 Variable 栏和 Minus 栏内输入要计算差值的变量，Minus 栏内是被减的变量，然后在 Percent 栏或 Arithmetic 栏内输入差值在报表中的标签。单击右边的箭头按钮移入 Pairs 框中，单击【Delete pair】按钮重新设置。

- “Differences between Groups of Cases”框

组对差值框，用于设定组对差值计算的参数。在“Grouping”栏内选择分组变量，按照这个变量分组对，在“Category”栏和“Minus”栏内输入计算差值的组对，“Minus”栏内是被减的组对，然后在“Percent”栏或“Arithmetic”栏内输入差值在报表中的标签。单击右边的箭头按钮移入“Pairs”框中，单击【Delete pair】按钮重新设置。

5. 【Title】子对话框

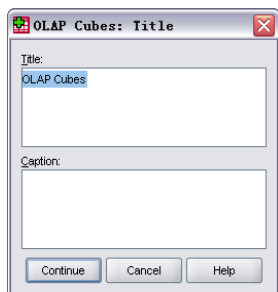


图 5-5 【Title】子对话框

【标题】子对话框。单击【Title】按钮，弹出如图 5-5 所示的子对话框，它主要用于定义报表的名称和说明。在“Title”框内输入报表的标题，在“Caption”框内输入报表的说明。

下面举例说明【OLAP】过程的使用法。

例 5.1 在数据文件“Employee Data.sav”中，以变量“jobcat”和“gender”为分组变量，以“salary”和“salbegin”为汇总变量，分析每组汇总变量均值和标准差，并计算各组中员工薪水的变化情况，将新表的标题设为“员工基本信息统计表”，并在说明部分写上制表时间和制表人。具体步骤如下：

执行【Analyze】/【Reports】/【OLAP Cubes】命令，弹出【OLAP Cubes】对话框	
Summary Variables : salary 和 salbegin	选择汇总变量
Grouping Variables : jobcat、gender	选择分组变量
单击【Statistics】按钮	弹出【Statistics】子对话框
Cell Statistics :	选择将要计算的统计量
Mean	
Standard Deviation	
单击【Continue】按钮	回到【OLAP Cubes】对话框
单击【Differences】按钮	弹出【Differences】子对话框
Differences for Summary Statistics 单选框 :	选择计算汇总变量之间的差值
“Differences between variables ”	
Type of Differences 复选框 : Arithmetic difference	选择差值类型
Differences between Variables 框内	进一步设置差值的参数
Variable : salary	
Minus variable : salbegin	
Arithmetic : 薪水变化水平	
单击【Continue】按钮	回到【OLAP Cubes】对话框
单击【Title】按钮	弹出【Title】子对话框
Title : 员工基本信息统计表	设置报表的标题和说明
Caption :	
制表时间 : 2009 年 10 月 20 日 制表人 : 杨紫	
单击【Continue】按钮	回到【OLAP Cubes】对话框
单击【OK】按钮	报表 5-2 和 5-3 出现在结果浏览窗口中

表 5-2 观测量汇总表

	Cases					
	Included		Excluded		Total	
	N	Percent	N	Percent	N	Percent
Current Salary* Employment Category*Gender	474	100.0%	0	.0%	474	100.0%
Beginning Salary* Employment Category*Gender	474	100.0%	0	.0%	474	100.0%

在结果浏览窗口中首先弹出的是如表 5-2 所示的观测量汇总表，表中列出了参与汇总分析的所有有效观测量数、被排除的观测量数，以及全部观测量数和它们所占的百分比。

表 5-3 员工基本信息统计表

Gender	Female	
Employment Category	Clerical	
	Mean	Std. Deviations
Current Salary	\$25,003.69	\$5,812.838
Beginning Salary	\$12,750.75	\$2,391.056
薪水变化水平	\$12,252.94	\$3,421.782

制表时间：2009年10月20日 制表人：杨紫

双击员工信息统计表得到表 5-3，可以看到“Gender”栏内和“Employment Category”栏都是可以通过选项调整的，单击后在报表中出现相应选项的统计值。同时，表格中新增了一项统计信息“薪水变化水平”。

5.1.2 观测量汇总——Case Summaries过程

对数据文件中的观测量进行汇总处理，就是【Case Summaries】过程所做的事情。

执行【Analyze】/【Reports】/【Case Summaries】命令，弹出如图 5-6 所示的【Summarize Cases】对话框，下面介绍其中元素。

1. 【Variables】框

变量框，其中放置将在表中列出的变量。

2. 【Grouping Variables】框

分组变量框，其中的变量是分组变量，将按其中的变量对观测量进行分组。

3. 【Display cases】复选框

显示观测量复选框，用于设定报表中将出现的观测量值的参数。一旦勾选【Display cases】选项，下面的 3 个选项都将被勾选。第一个选项“Limit cases to first”，后面的空白栏里指定观测量（默认值是 100），输入一个指定值，它是指在报表中显示数据文件从第一个观测量开始到指定值的数目；第二个选项“Show only valid cases”，指只显示有效观测值，

缺失值排除在外；第三个选项“Show case numbers”，指将被选中的观测量在数据文件中的序号也显示在表中。

4. 【Statistics】子对话框

【统计量】子对话框。单击【Statistics】按钮弹出类似于图 5-3 所示的子对话框。两者的区别只是系统默认值不同，这里的系统默认值只有统计量“Number of cases”。

5. 【Options】子对话框

【选项】子对话框。单击【Options】按钮，弹出如图 5-7 所示的子对话框，类似于 Title 子对话框，它除了用于定义报表的名称和说明外，还用于设定一些报表显示的细节。相同的部分这里就不再介绍了，主要介绍以下三个选项：

- Subheadings for totals，选择对每个分组是否显示所计算的统计量的名称；
- Exclude cases with missing value listwise，选择在分析中是否删除带有缺失值的变量；
- Missing statistics appear as 栏，输入某个字符、文字或者短语，用来标记数据文件中的缺失值。

注意 若选择此项，则不能选择主对话框中的“Show only valid case”选项。

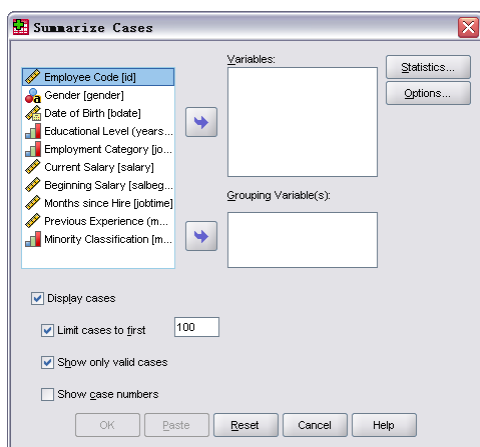


图 5-6 【Summarize Cases】对话框

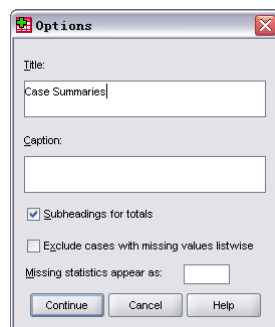


图 5-7 【Options】子对话框

下面举例说明【Case Summaries】过程的用法。

例 5.2 要求利用文件“Employee Data.sav”的前 12 个观测量值制作一个“员工信息汇总表”，按照变量 gender 对数据进行分组，对每组及所有的变量“salary”、“prevexp”和“jobcat”计算它们的观测量数目、均值和极值，并且不排除缺失值。具体操作步骤如下：

执行【Analyze】/【Reports】/【Case Summaries】命令，弹出【Summarize Cases】对话框

Variables : jobcat、prevexp 和 salary

选择汇总变量

Grouping Variables : gender

选择分组变量

勾选“Display cases”

选择将在报表中显示的观测量

Limit cases to first : 12

限制观测量数目

去除勾选 “ Show only valid cases ”	缺失值也要显示在报表中
去除勾选 “ Show case numbers ”	不用显示观测量在数据文件中的序列号
单击【Statistics】按钮	弹出【Statistics】子对话框
Cell Statistics : Number of Cases	选择将要计算的统计量
Mean	
Range	
单击【Continue】按钮	回到【Summarize Cases】对话框
单击【Options】按钮	进入【Options】子对话框
Title : 员工信息汇总表	设置报表的标题
勾选 “ Subheadings for totals ”	在表中的分组内显示 Total
单击【Continue】按钮	回到【Summarize Cases】对话框
单击【OK】按钮	报表出现在结果浏览窗口，如表 5-4 所示

同 OLAP 过程一样，在这个报表的前面还有一个观测量汇总表，与表 5-2 类似，在其中列出了参与汇总分析的所有有效观测量数、被排除的观测量数，以及全部观测量数和它们所占的百分比。

表 5-4 员工信息汇总表

			当前工资	工作经验（月）	职 业
性别 女	1		\$21 450	381	文书
	2		\$21 900	190	文书
	3		\$21 900	missing	文书
	4		\$27 900	115	文书
	5		\$24 000	244	文书
	6		\$30 300	143	文书
	汇总	N	6	6	6
		Range	\$8 850	381	0
		Mean	\$24 575.00	178.83	1.00
男	1		\$57 000	144	经理
	2		\$40 200	36	文书
	3		\$45 000	138	文书
	4		\$32 100	67	文书
	5		\$36 000	114	文书
	6		\$28 350	26	文书
	汇总	N	6	6	6
		Range	\$28 650	118	2
		Mean	\$39 775.00	87.50	1.33
汇总	N		12	12	12
	Range		\$35 550	381	2
	Mean		\$32 175.00	133.17	1.17

a. Limited to first 12 cases.

5.1.3 生成商务报表——Report Summaries in Rows/Columns过程

【Report Summaries in Rows】与【Report Summaries in Columns】是 SPSS 的两个相似的功能，它们的区别仅在于输出方式的不同，分别指按行输出和按列输出。与观测量汇总过程相比，这两个过程的功能更加强大。

1. Report Summaries in Rows

执行【Analyze】/【Reports】/【Report Summaries in Rows】命令，弹出如图 5-8 所示的【Summaries in Rows】对话框。

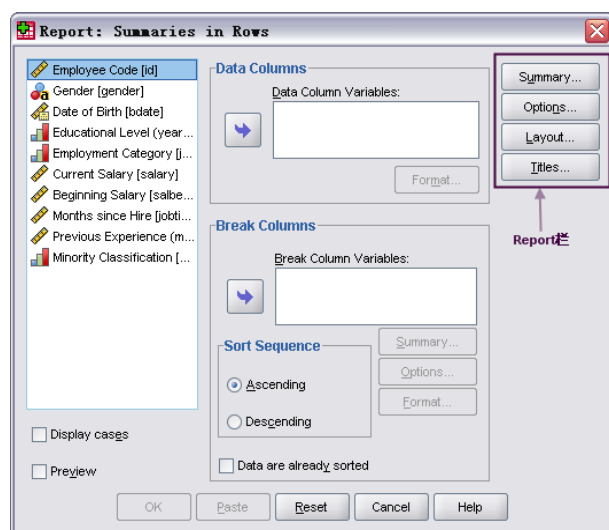


图 5-8 【Summaries in Rows】对话框

- 【Data Columns】框

数据列框，用于设置报表中的数据列。将需要进行汇总的变量从原变量列表框移到【Data Columns】框中。

- 【Break Columns】框

分组列框，用于设置报表中的分组列，即报表中的第一列。将分组变量移入这个框中。

- 【Sort Sequence】框

排序框，用于确定分组变量输出时的顺序。

- “Data are already sorted” 选项

数据已经经过排序选项。当数据文件已经按照分组变量排好序了，可以选择此选项，这时，系统不再对数据进行分类排序。

- “Display cases” 选项

显示观测量值选项，若选择此项，文件中各分组的所有观测量都将列在报告中。

- “Preview” 选项

预览选项，若选择此项，系统将产生一页预览表。

• 【Format】子对话框

格式子对话框，用于定义变量在报表中出现的格式。在【Data Columns】框和【Break Columns】框下面都有一个【Format】按钮，选中这两个框中的任意一个变量后，单击这个变量所对应的【Format】按钮，弹出如图 5-9 所示的子对话框（这里选中的变量是“salary”）。

① Column Title 框：列标题框，在其中输入变量“salary”在报表中的列标题。

② Column title justification 矩形框：列标题对齐方式矩形框，其中有三个选项，分别是 Left（左对齐）、Center（居中）和 Right（右对齐）。

③ Column Width 栏：列宽栏，在其中输入一个数值，用于指定列宽。

④ Value Position within Column 单选框：列中变量位置单选框，用于指定列中变量的位置。第一个选项 Offset from right（从右开始缩进），在“Offset amount”栏内输入一个数值指定缩进量，如果选入的变量是字符型的，则这个选项就变成 Offset from left（从左开始缩进）；第二个选项“Centered within column”指将变量值或列标签位于列中央。

⑤ Column Content 单选框：列内容单选框，第一个选项 Values 意思是输出中显示变量值（系统默认选项）；第二项 Values labels 意思是输出中显示变量的标签值。

• 【Summary Lines】子对话框

对分组变量按行汇总对话框，在其中选择分组后每组变量的汇总统计量。单击【Break Columns】框下面的【Summary】按钮，弹出如图 5-10 所示对话框。其中的所有统计量意义在表 5-1 中都可以找到。

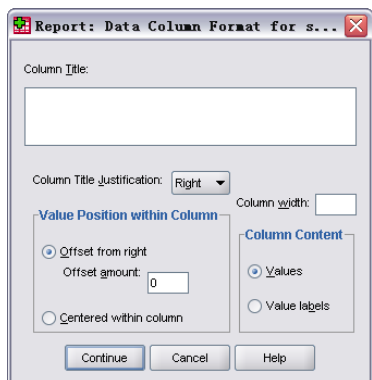


图 5-9 【Data Column Format】子对话框

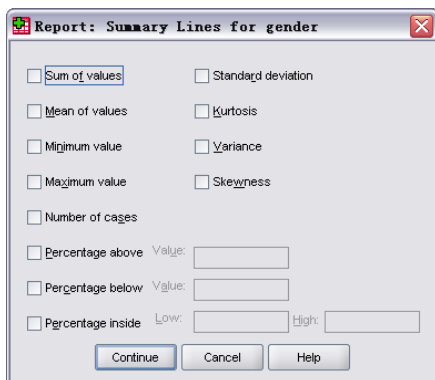


图 5-10 【Summary Lines】子对话框

• 【Break Options】子对话框

分组变量选项对话框，用于指定页面控制。单击【Break Columns】框下面的【Options】按钮，弹出此对话框，如图 5-11 所示。

① Page Control（页面控制）框：第一个选项“Skip lines before break”是指在各分组之间插入指定数目的空行，这个指定数目输入到此选项后面的空白栏内（只能是 0~20 之间的数值）；第二个选项“Begin next page”，指按每组一页，页码连续的格式输出报告；第三个选项“Begin new page & reset page number”，指按每组一页，在新的一页开始时列置汇总变量，并重置页码的格式输出报告。

② **Blank Lines before Summaries** 栏: 汇总值前空白行数栏, 在其后的空白栏内输入 0~20 之间的一个数值, 输出时在各分组的标签值与汇总报告之间插入指定数目的空行。

• **【Report】** 框

在图 5-8 所示的对话框中, 最右边的 4 个按钮统称为报表框, 用于对数据文件中全部数据的总汇总结果进行修饰与控制。上面有 4 个选项按钮。

① **【Summary】** 子对话框: 汇总子对话框, 用于选择对所有观测量进行汇总统计的统计量。这个对话框与图 5-10 所示类似。

② **【Options】** 子对话框: 选项子对话框, 用于确定缺失值处理方式及页码设置。单击 **【Report】** 框中的 **【Options】** 按钮, 弹出此对话框, 如图 5-12 所示。

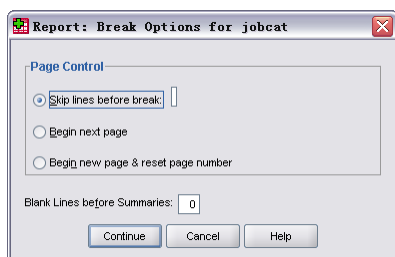


图 5-11 **【Break Options】** 子对话框

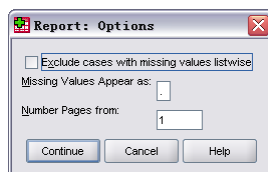


图 5-12 **【Options】** 子对话框

Exclude cases with missing values listwise: 将有缺失值的观测量全部排除。

Missing values Appear as: 缺失值显示为指定符号, 这个指定符号输入到后面的空白栏内, 系统默认的符号是句点 “.”。

Number Pages from 栏: 首页页码设置栏, 用于设定首页的页码, 系统默认值是 “1”。

③ **【Layout】** 子对话框: 布局子对话框, 用于设置报告的布局, 如图 5-13 所示。其中 “**Page Layout**” 框用于设置页面的布局, 包括设置报告页的起始行数、结束行数, 页码的左页边距、右页边距, 以及页码对齐方式; “**Page Titles and Footers**” 框用于设置页标题和脚注, 包括设置标题和报告首行之间的空行数、报告末行和脚注之间的空行数; “**Column Titles**” 框用于设置列标题, 包括设置列标题下是否加下画线, 列标题与报告首行之间的空行数, 列标题的对齐方式; “**Break Column**” 框用于有多个分组变量时设置所有分组变量的排列方式; “**Data Column Rows & Break Labels**” 框用于设置第一个统计量与分组变量之间的位置关系。

④ **【Titles】** 子对话框: 标题子对话框, 用于设置报告的标题和脚注。单击 **【Titles】** 按钮, 弹出此子对话框, 如图 5-14 所示。在此对话框右侧的上下栏内可以为报告设置多达 10 行的标题和脚注, 可以直接输入, 也可以从原变量列表框中选择变量作为标题和脚注。这里特别指出, 左下角 “**Special Variables**” 框内的两个特殊变量 **DATA** (日期) 和 **PAGE** (页码), 将它们选入右面的标题栏或者脚注栏内就会自动在报告的相应位置显示当日日期和页码了。其中的 **【Previous】** 按钮和 **【Next】** 按钮用于翻页。

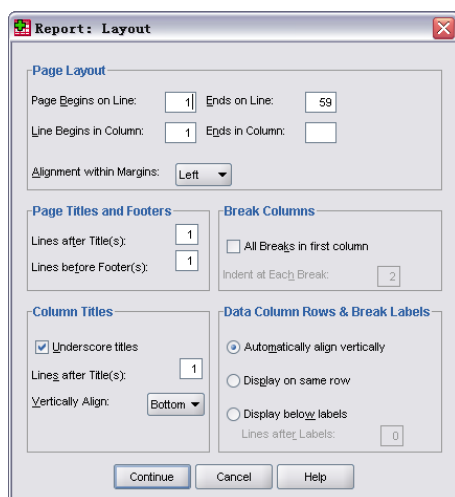


图 5-13 【Layout】子对话框

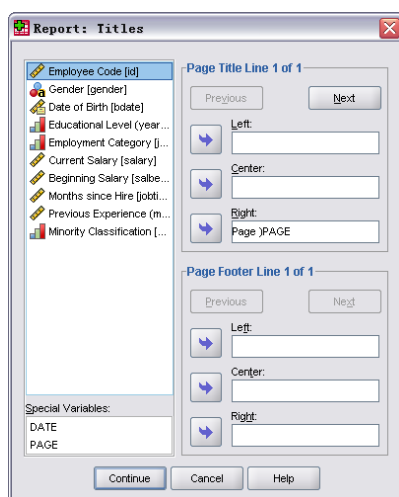


图 5-14 【Titles】子对话框

下面举例说明【Report Summaries in Rows】过程的使用法。

例 5.3 在数据文件“Employee Data.sav”中，要求以变量“jobcat”为分组变量，以变量“salary”和“jobtime”为数据列变量，设置列宽为 15，对各个分组及所有观测量统计它们的均值、最小值、最大值和观测量数目，报告的标题设为“各工种及全体员工工资和工龄报告”，并在报告上右方记录报告的日期，其他的格式按照系统默认值设置。具体操作步骤如下：

执行【Analyze】/【Reports】/【Case Summaries】命令，弹出【Summaries in Rows】对话框

Data Columns : salary	选择数据列变量 “ salary ”
单击【Format】按钮	弹出【Format】子对话框
Column Title : 工资	设置变量 “ salary ” 在报告中的格式
Column Width : 15	
单击【Continue】按钮	回到主对话框
Data Columns : jobtime	选择数据列变量 “ jobtime ”
单击【Format】按钮	弹出【Format】子对话框
Column Title : 工龄	设置变量 “ jobtime ” 在报告中的格式
Column Width : 15	
单击【Continue】按钮	回到主对话框
Break Columns : jobcat	选择分组变量 “ jobcat ”
单击【Summary】按钮	弹出【Summary Lines】子对话框
勾选 Mean of values、Minimum value Maximum value、Number of cases	选择各分组统计量
单击【Continue】按钮	回到主对话框
单击【Format】按钮	弹出【Format】子对话框
Column Title : 工种	设置变量 “ jobcat ” 在报告中的格式
Column Width : 15	

单击【Continue】按钮	回到主对话框
单击“Report”框中的【Summary】按钮	弹出【Summary】子对话框
勾选 Mean of values、Minimum value Maximum value、Number of cases	选择整体的统计量
单击【Continue】按钮	回到主对话框
单击“Report”框中【Title】按钮	弹出【Title】子对话框
Page Title Line 1 of 2 : Center : 各工种及全体员工工资和工龄报告 Page Title Line 2 of 2 : Right : PAGE	设置报告标题
单击【Continue】按钮	回到主对话框
单击【OK】按钮	表 5-5 出现在结果浏览窗口中

2. Report Summaries in Columns

执行【Analyze】/【Reports】/【Summaries in Columns】命令，弹出如图 5-15 所示的【Summaries in Columns】对话框。这个对话框与【Summaries in Rows】对话框类似，差别是两者的【Summary】按钮位置不同，两个【Option】子对话框的内容有部分增加。

表 5-5 按行汇总报告实例

各工种及全体员工工资和工龄报告		
07 Nov 06		
工 种	工 资	工 龄
Clerical		
Mean	\$27 839	81
Minimum	\$15 750	63
Maximum	\$80 000	98
N	363	363
Custodial		
Mean	\$30 939	82
Minimum	\$24 300	67
Maximum	\$35 250	95
N	27	27
Manager		
Mean	\$63 978	81
Minimum	\$34 410	64
Maximum	\$135 000	98
N	84	84
Grand Total		
Mean	\$34 420	81
Minimum	\$15 750	63
Maximum	\$135 000	98
N	474	474

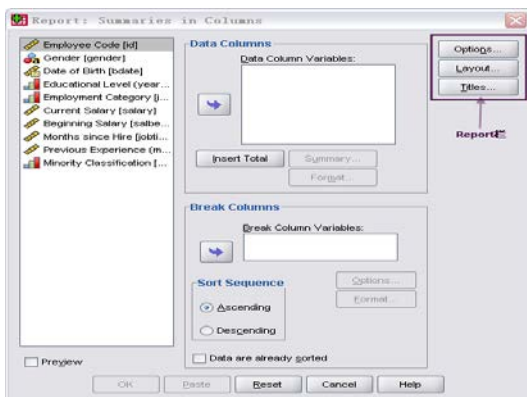


图 5-15 【Summaries in columns】对话框

- 【Summary Line】子对话框

汇总子对话框,用于定义每个分组的一个汇总统计量。选入一个变量到【Data Columns】框中,单击下面的【Summary】按钮,弹出此对话框,如图 5-16 所示。注意它与【Summaries in Rows】对话框中的【Summary Line】子对话框最大的不同在于它是一个单选框,而后者是复选框。

- 【Summary】子对话框

总汇总子对话框,用于定义每个分组中一个总的汇总统计量。当所有的数据列变量及其统计量选择好以后,单击【Insert Total】按钮,然后再单击【Summary】按钮,弹出此对话框,如图 5-17 所示。左侧的 Data Columns 框内列出了所有被选好的数据列变量及其统计量,将所需的数据列变量及其统计量移入“Summary Column”框中,然后在“Summary function”下拉列表框中选择一个统计量即可。

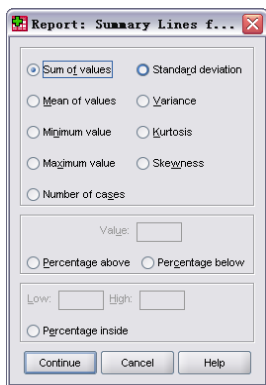


图 5-16 【Summary Lines】子对话框

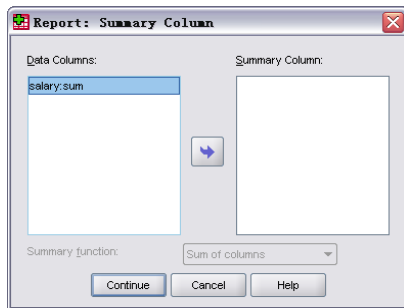


图 5-17 【Summary Column】子对话框

- 【Break Options】子对话框

分组变量选项对话框,用于指定页面控制。单击【Break Columns】框下面的【Options】按钮,弹出此对话框,如图 5-18 所示。这个对话框与【Summaries in Rows】对话框中的【Break Options】子对话框相比,多了一个“Subtotal”(分组小计)框,若选择其中的选项

“Display subtotal”，则报告将显示每个分组的小计，此时，Label 栏被激活，在下面输入分组变量的标签。

• 【Options】子对话框

选项子对话框，用于确定缺失值处理方式及页码设置。单击【Report】框中的【Options】按钮，弹出该对话框，如图 5-19 所示。这个对话框与【Summaries in Rows】对话框中的【Options】子对话框相比，多了一个“Grand Total”（总计）框，若选择其中的选项“Display grand total”，则报告将显示总汇总表，此时，“Label”栏被激活，在下面输入总计标签。

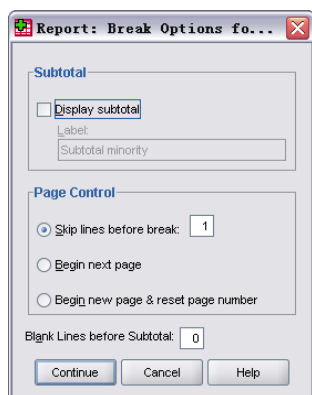


图 5-18 【Break Options】子对话框

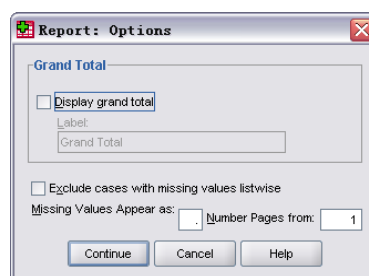


图 5-19 【Options】子对话框

下面举例说明【Report Summaries in Columns】过程的用法。

例 5.4 要求以变量“jobcat”为分组变量，统计不同工种工资情况，其中的数据列包括工资的最小值、最大值、工资大于\$30 000 美元的人数占每个组人数的百分比，及工资的均值，并总结每个分组中工资最大值与最小值的差值。

具体操作步骤如下：

执行【Analyze】/【Reports】/【Summaries in Columns】命令，弹出【Summaries in Rows】对话框	
Data Columns：salary	选择数据列变量“salary”
单击【Summary】按钮	弹出【Summary Lines】子对话框
选择 Minimum value	选择统计量：最小值
单击【Continue】按钮	回到主对话框
单击【Format】按钮	弹出【Format】子对话框
Column Title：工资	设置变量“salary”在报告中的格式
Column Width：12	
单击【Continue】按钮	回到主对话框
用同样的方法设置工资的最大值、工资大于 30 000 美元的人数占每个组人数的百分比、工资的均值	
单击【Insert Total】按钮	选入一系列总结量
单击【Summary】按钮	弹出【Summary】子对话框

Summary Column : salary : max	总结每个分组中工资最大值与最小值的差值
Salary : min	
Summary function : 1st column-2 nd column	
单击【Continue】按钮	回到主对话框
Break Columns : jobcat	选择分组变量 “ jobcat ”
单击【Format】按钮	弹出【Format】子对话框
Column Title : 工种	设置变量 “ jobcat ” 在报告中的格式
Column Width : 15	
单击【Continue】按钮	回到主对话框
单击 “ Report ” 框中【Title】按钮	弹出【Title】子对话框
Page Title Line 1 of 1 :	
Center : 不同工种工资情况报告	设置报告标题
Right : Page]PAGE	
单击【Continue】按钮	回到主对话框
单击【OK】按钮	表 5-6 出现在结果浏览窗口中

表 5-6 按列汇总报告实例

不同工种工资情况报告			Page	1	
工 种	工资最小值	工资最大值	工资>\$30 000	工资均值	Total
Clerical	\$15 750	\$80 000	29.5%	\$27 839	\$64 250
Custodial	\$24 300	\$35 250	74.1%	\$30 939	\$10 950
Manager	\$34 410	\$135 000	100.0%	\$63 978	\$100 590

5.2 高级报表——Tables子菜单

关于报表的另一个菜单就是【Analyze】菜单下的【Tables】子菜单。它主要是为生成高级报表而设计的，功能非常强大，甚至还允许用户自定义报表。该菜单如图 5-20 所示。



图 5-20 【Tables】子菜单

5.2.1 定义复选变量集——Multiple Response Sets过程

执行【Analyze】/【Tables】/【Multiple Response Sets】命令，弹出如图 5-21 所示的对话框，这个命令实际上就是【Data】/【Define Multiple Response Sets】命令，它用于为当前数据文件定义复选变量集，该命令在这里出现是因为复选变量在高级报表中有着非常重要的作用。下面举一个例子来说明复选变量的意义。

例 5.5 某地开展了一项关于当地人口休闲活动的调查，调查内容包括性别、年龄、月收入、学历、休闲时间，以及平时主要的休闲活动是哪些，答案包括旅游休闲、看电视、做运动、做游戏，以及其他休闲活动（此项为多选）。将每一个选项作为一个变量输入到名为“某地休闲活动调查.sav”的数据文件中（具体数据见光盘中），变量值为 1 时表示该选项被选中。

下面来介绍如图 5-21 所示的【Define Multiple Response Sets】对话框中的主要元素。

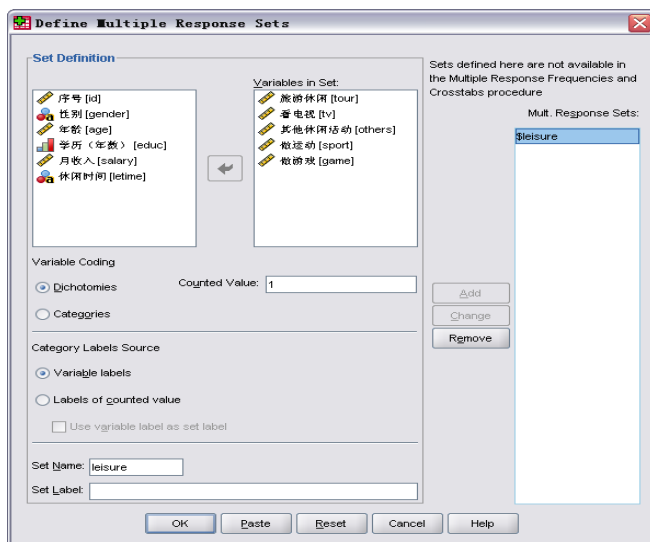


图 5-21 【Define Multiple Response Sets】对话框

1. 【Mult. Response Sets】框

复选变量集框，放置已经定义好了的复选变量。

2. 【Variables in Set】框

变量集框，用于放置一个复选变量中的所有变量。

3. 【Variable Coding】单选框

变量编码方式单选框，用于选择变量集中变量的编码方式。其中包含两个选项。

• Dichotomies 选项

多重二分法。这种分类方法适合多选题的选项数目不大时使用，比如，本例中的多选题，将每个选项作为一个变量，以数值 1 表示该选项被选中，也就是说每个变量都是一个有 2 个分类的分类变量。后面的“Counted Value”栏内输入表示该选项被选中的数值。

• Categories 选项

多重分类法。这种分类法适合多选题的选择数目较大时使用，例如，在一次调查活动中需要被调查者选择自己最喜爱的 10 个商业广告，其中待选项有 70 个。这时，若选用上面的方法就需要在数据文件中建立 70 个变量，显然是很麻烦的。因此换一种方法，将被调查者的每一次选择作为变量，变量值就是被选中的选项内容。这样，文件中就只需要建立 10 个变量，每个变量都是有多个分类的分类变量，这就是多重分类法。

4. 【Category Labels Source】单选框

类型标签源单选框，用于选择新复选变量中每一个分类的标签是来自原变量的标签（第一个选项），还是来自被选中的变量值的标签（第二个选项）。

5. 【Set Name】和【Set Label】栏

设置复选变量名称和标签栏。在其后的空白栏内分别输入新复选变量的名称和标签。上面有个选项 Use variable label as set label，如果选中，变量的标签就直接赋给复选变量。

若要求将变量“tour”、“tv”、“sport”、“game”和“others”定义成一个复选变量，令其名称为“leisure”，标签为“休闲活动种类”。具体操作步骤如下：

Variables in Set : tour、tv、sport、game、others	将包含在变量集中的 5 个变量选入
Variable Coding : Dichotomies	选择多重二分法
Counted Value : 1	变量取值为 1 时表示该答案被选中
Set Name : leisure	定义复选变量集名称和标签
Set Label : 休闲活动种类	
单击【Add】按钮	添加所定义的复选变量集
单击【OK】按钮	复选变量自动被数据文件保存
结果浏览窗口中出现所定义的复选变量集的性质报表，如表 5-7 所示。	

表 5-7 复选变量集报表

Multiple Response Sets				
Name	Coded AS	Counted Value	Data Type	Elementary Variables
\$leisure	Dichotomies	1	Numeric	旅游休闲 看电视 其他休闲活动 做运动 做游戏

5.2.2 定制报表——Custom Tables过程

SPSS【Custom Tables】过程功能大大超越了上面提到的其他报表功能，它采用所见即所得的方式，极大地减少了对话框中数目多且烦杂的子对话框和选项，易学易用，并用户可以随心所欲地定制报告的格式和内容。其他的每一个报表功能都具有自己的特点，用于满足某个特定的需要，但是【Custom Tables】功能强大到拥有其他所有报表功能的特点，其他报表功能所能生成的报表，都可以通过【Custom Tables】功能来生成。它还可以根据生成的报表进行卡方检验、T 检验和 Z 检验。

执行【Analyze】/【Tables】/【Custom Tables】命令，弹出如图 5-22 所示的对话框，其中有 4 个选项卡，【Table】选项卡为系统默认选项。下面介绍每个选项卡中的元素。

1. 【Table】选项卡

报表选项卡，用于设置报表的内容。

• 画布框

中间的大空白框，用于显示报表的内容和排列方式。其中包含一个 Column 图标，一个 Rows 图标，将变量拖入 Rows 图标，当图标外框显示成红色时，表示该变量可以被选为行变量，同样的方法可以选择列变量。上方有 2 个按钮，系统默认为【Normal】状态，是指显示标准的报表样式，单击【Compact】按钮，显示报表的简洁格式。

• 【Layers】框

分层汇总变量框，用于设置汇总变量。单击【Layers】按钮，才会出现这个框，上面的框中显示被选入的分层变量，当分层变量的数目多于一个时，下面的【Layer Output】单选框将被激活，第一个选项的意思是分别显示每个分类的分层，第二个选项的意思是显示每个分类的组合。

• 【Category Position】框

分类变量位置框，用于选择分类变量的位置。当画布框中只有一个分类变量时，该框才被激活，default 选项是指使用当前的显示方式，另一个选项的意思是转换当前位置，即如果原来是行变量就转换为列变量，如果原来是列变量则转换为行变量。

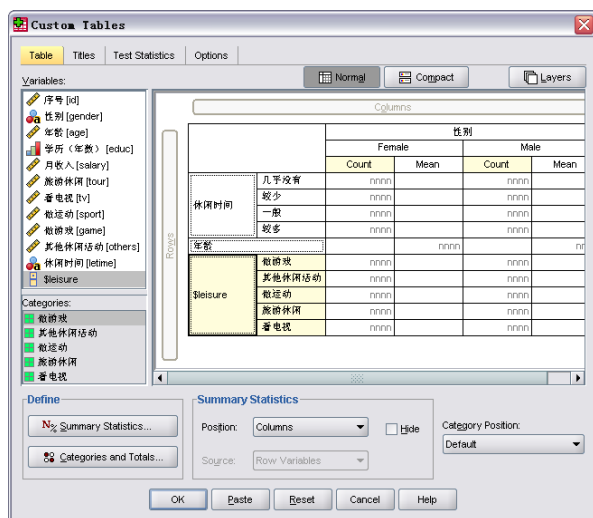


图 5-22 【Custom Tables】对话框的 Table 选项卡

• 【Summary Statistics】框

汇总统计框，用于设置汇总统计量显示的格式。【Position】栏内选择汇总统计量所在的位置是在行中还是在列中，Hide 选项的意思是隐藏统计量标签。【Source】栏内选择统计量的来源。

• 【Categories and Totals】子对话框

分类与汇总子对话框。在选中画布框中的任意分类变量后，单击【Categories and Totals】按钮，弹出如图 5-23 所示的子对话框（这里选中的是变量“sex”），Display 框中显示的是

变量的所有分类，在 New Subtotal 框内有一个【Insert】按钮，可以为分类变量添加新的分类，在 Label 栏内输入这个分类的标签，右面的【Delete Subtotal】按钮用于删除变量中的分类，被删除的分类被移入 Exclude 框内；Sort Categories 框用于为分类排序；Show 框用于选择在分类变量栏内显示的内容，其中包括 Total（汇总）、Missing Values（缺失值）、Empty categories（空白分类）和 Other values found when data are scanned（对数据检测时发现的其他变量值），勾选其中的选项表示显示这个选项所指的内容。

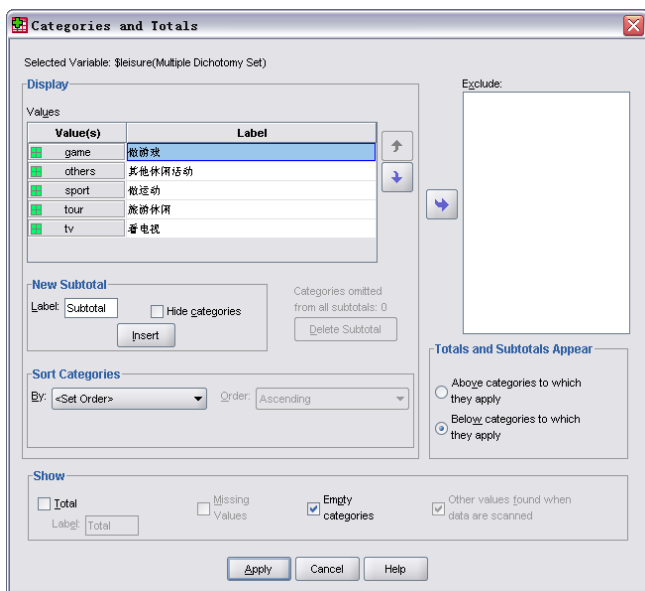


图 5-23 【Categories and Totals】子对话框

• 【Summary Statistics】子对话框

汇总统计对话框。若选中变量是 Scale 型变量，单击【Summary Statistics】按钮，弹出如图 5-24 所示的子对话框，若选中 Categorical 变量(分类变量)，单击【Summary Statistics】按钮，弹出如图 5-25 所示的子对话框。它们的 Display 框样式是相同的，在“Statistics”栏内显示统计量的类型，在“Label”栏内显示并可以修改统计量的标签，在“Format”栏内显示统计数据的输出格式，在“Decimal”栏内显示统计数据的小数部分位数。在图 5-25 中可以看到有两个统计量列表框，下面的列表框是专为分类变量的汇总定制统计量的，勾选前面的选项表示要定制汇总统计量。

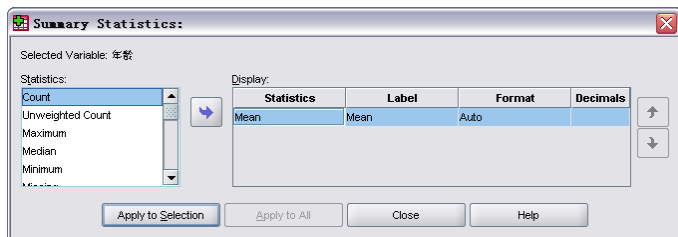


图 5-24 Scale 型变量的【Summary Statistics】子对话框

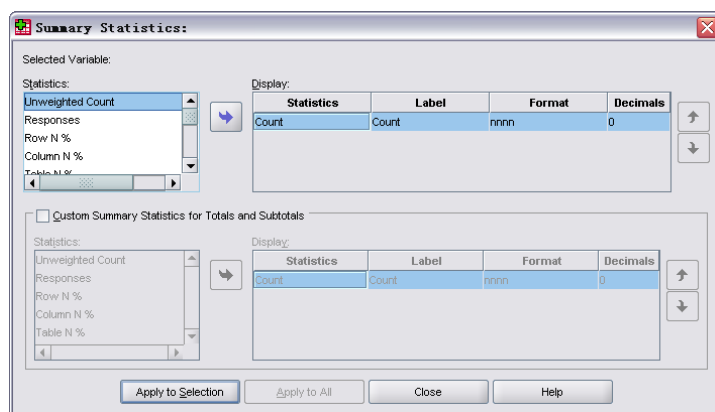


图 5-25 分组变量的【Summary Statistics】子对话框

2. 【Title】选项卡

标题选项卡，用于设置报表的标题。与上一章看到过的【Title】子对话框类似，不过多了一个【Corner】框，用于设置报表最左上角单元格中显示的文本。

3. 【Test Statistics】选项卡

统计检验选项卡，用于设置对报表中数据进行统计检验的参数。其中包括卡方检验、T 检验和 Z 检验。

4. 【Options】选项卡

选项选项卡，用于设置报表的一些基本格式，如图 5-26 所示。

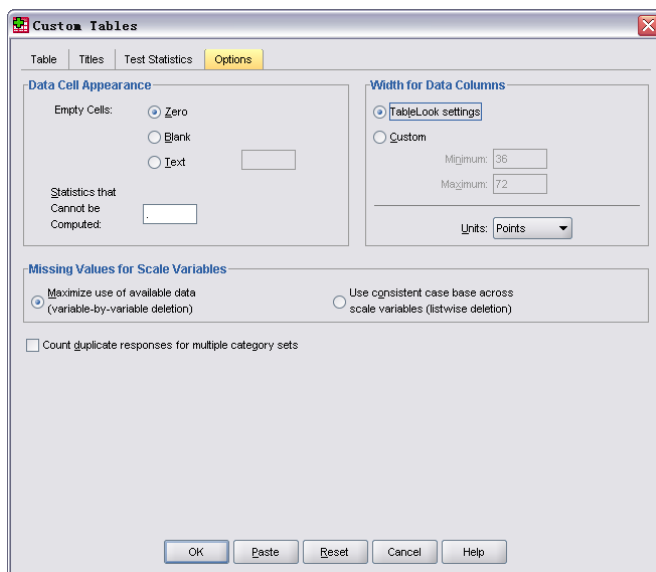


图 5-26 【Custom Tables】对话框的 Options 选项卡

- **【Data Cell Appearance】** 框

数据单元格显示框，用于设置空单元格的显示方式。空单元格的显示方式包括零 (Zero)、空白 (Blank) 和文本 (Text)，在后面的空白栏内输入文本内容；在 Statistics that Cannot be Computed 栏内输入无法计算统计量的栏内的输出内容，系统默认为 “.”。

- **【Width for Data Columns】** 框

数据列宽度，用于设置数据列的宽度。系统默认为第一个选项，第二个选项意思是让用户自定义，在后面的 Minimum 栏和 Maximum 栏内分别输入最小和最大长度，Units 栏选择所输入长度的单位。

- **【Missing Values for Scale Variables】** 框

Scale 型变量的缺失值框，用于设置 Scale 型变量的缺失值处理方式。

例 5.6 对数据文件“某地休闲活动调查.sav”进行分类，建立一个报表，按照不同性别描述变量“letime”、“age”和“\$leisure”，还要对变量“sex”进行汇总。其中，要求输出休闲时间的频数，输出年龄的均值，还要求分别输出复选变量中的 5 个变量的频数和占总例数的百分比。报表取名为“某地休闲活动调查报表”，在报表下面显示报表生成的时间。具体操作步骤如下。

进入【Table】选项卡：	
Columns 图标：sex	将列变量性别拖曳到画布框
Position：Rows	将统计量标签移到行变量后面
单击【Categories and Total】按钮	进入子对话框
勾选 Total	将所有性别汇总
去掉勾选 Show 框中最后一个选项	
单击【Apply】按钮	回到主对话框
Rows 图标：letime	将行变量休闲时间拖曳到画布框
单击【Categories and Total】按钮	进入子对话框
去掉勾选 Show 框中最后一个选项	
单击【Apply】按钮	回到主对话框
Rows 图标：age	将行变量年龄拖曳到画布框
Rows 图标：\$leisure	将行变量休闲活动种类拖曳到报表显示框
单击【Summary Statistics】按钮	进入子对话框
Display：Table Response%	计算总体比例
Label：总体比例	更改 Table Response% 的输出标签
单击【Apply to Selection】按钮	回到主对话框
切换到“Title”选项卡	
Title：某地休闲活动调查报表	输入报表名称
Caption：Date	输入报表生成的日期
单击【OK】按钮	生成的报表出现在结果浏览窗口中 如表 5-8 所示

表 5-8 自定义报表

某地休闲活动调查报表

			性 别		
			Female	Male	Total
休闲时间	几乎没有	Count	1	0	1
	较少	Count	12	17	29
	一般	Count	11	16	27
	较多	Count	9	4	13
年龄	Mean		40	40	40
Sleisure	旅游休闲	Count	8	3	11
		总体比例	8.2%	3.1%	11.2%
	看电视	Count	15	17	32
		总体比例	15.3%	17.3%	32.7%
	其他休闲活动	Count	5	8	13
	活动	总体比例	5.1%	8.2%	13.3%
	做运动	Count	10	10	20
		总体比例	10.2%	10.2%	20.4%
	做游戏	Count	9	13	22
		总体比例	9.2%	13.3%	22.4%

2009-10-15

5.3 本章小结

本章介绍了 SPSS 报表。SPSS 报表分为简单记录报表【Reports】子菜单和高级报表【Tables】子菜单两类，其中需要重点掌握的是【Tables】子菜单，尤其是其定义复选变量集【Multiple Response Sets】过程。同时，掌握利用这两个子菜单下的其余过程制作统计报表。

PART

第 3 篇 统计分析

- 第 6 章 描述性统计分析
- 第 7 章 均值比较与 t 检验
- 第 8 章 方差分析
- 第 9 章 相关分析
- 第 10 章 回归分析
- 第 11 章 聚类分析与判别分析
- 第 12 章 因子分析与对应分析
- 第 13 章 非参数检验

第 6 章 描述性统计分析

前面几章都是在为统计分析做准备。从本章开始，我们将正式进入统计分析的学习。为了使用户能够正确运用恰当的统计方法并且对 SPSS 的输出结果给出合理的解释，对于所涉及的统计学概念和方法，尽可能用精练的语言给出其解释。本章的内容包括：

- 描述性统计量
- 频数分布表分析——Frequencies 过程
- 最基础的统计量分析——Descriptive 过程
- 探索性分析——Explore 过程
- 列联表分析——Crosstabs 过程
- 相对比描述——Ratio 过程

6.1 描述性统计量

描述性统计分析是基础的统计分析过程。对于整理好的数据，通过描述性统计分析，可以挖掘出很多统计量的特征。本节首先介绍在 SPSS 描述性统计分析过程中经常出现的各类描述性统计量。

6.1.1 描述性统计量

顾名思义，描述性统计量是指描述变量某一特征的统计量。常见的描述性统计量主要包括以下三类：描述变量集中趋势的统计量、描述变量离散程度的统计量、描述变量分布情况的统计量。表 6-1～表 6-3 分别介绍了 SPSS 中经常出现的这三类统计量。

表 6-1 描述变量集中趋势的统计量

中 文 名	SPSS 中表示	数学含义	特 点
均值	Mean	表示某变量的所有变量值的集中趋势或平均水平	容易受极值或异常值干扰
中位数	Median	一组数据中恰好使累积概率取 1/2 的变量值	较稳定，不易受极值或异常值干扰
众数	Mode	一组数据中出现频数最多的变量值	可能不唯一，多用于定类变量。不易受极值或异常值干扰
和	Sum	表示某变量的所有变量值的和	——

表 6-2 描述变量离散程度的统计量

中 文 名	SPSS 中表示	数学含义	SPSS 中计算公式
标准差	Std.deviation	描述变量关于均值的扰动程度	$\sqrt{\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2}$
方差	Variance	标准差的平方	$S = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2$
极小值	Minimum	某变量所有取值的最小值	——
极大值	Maximum	某变量所有取值的最大值	——
全距	Range	某变量极大值与极小值之差	Maximum – Minimum
均值的标准误差	S.E.mean	均值的标准误差	$SE = \frac{S}{\sqrt{n}}$

表 6-3 描述变量分布情况的统计量

中文名	SPSS 中表示	数学含义	SPSS 中计算公式
偏度	Skewness	描述某变量分布的对称程度和方向。偏度为 0 表示对称，大于 0 表示右偏，小于 0 表示左偏	$Skewness = \frac{\sum_{i=1}^n (x_i - \bar{x})^3}{(n-1)(n-2)S^3}$
峰度	Kurtosis	描述某变量分布的陡峭程度。峰度为 0 表示陡峭程度与正态分布相同，大于 0 表示比正态分布陡峭，小于 0 表示比正态分布平缓	$Kurtosis = \frac{n(n+1) \sum_{i=1}^n (x_i - \bar{x})^4 - 3(n-1) \sum_{i=1}^n (x_i - \bar{x})^2 \sum_{i=1}^n (x_i - \bar{x})^2}{(n-1)(n-2)(n-3)S^4}$

6.1.2 Descriptive Statistics子菜单概述

如图 6-1 所示，【Descriptive Statistics】子菜单共包括 7 个过程。各个过程的主要作用如下。

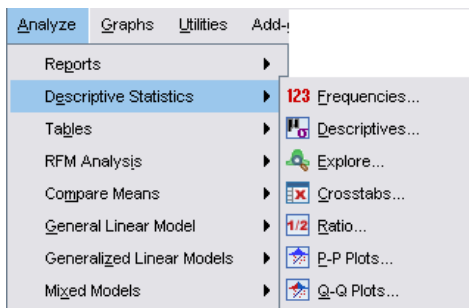


图 6-1 【Descriptive Statistics】子菜单

- ① **Frequencies**: 产生变量值的频数分布表，并可计算常见描述性统计量和绘制相对应的统计图。
- ② **Descriptives**: 计算一般的描述性统计量。
- ③ **Explore**: 探索性分析，使用户能够从大量的分析结果之中挖掘到所需要的统计信息。

④ Crosstabs: 对分类变量进行统计推断, 包括 χ^2 检验、确切概率等, 是 SPSS 重要的过程。

⑤ Ratio: 计算两个变量相对比的统计量特征。

⑥ P-P Plots: 绘制 P-P 图, 检验数据服从的分布情况。

⑦ Q-Q Plots: 绘制 Q-Q 图, 检验数据服从的分布情况。

其中, 【P-P Plots】和【Q-Q Plots】这两类过程已经在第4章介绍过了。在本章的后面几节, 将通过具体的例子来学习余下5类过程。

6.2 频数分布表分析——Frequencies过程

在介绍【Frequencies】过程之前, 首先强调一点, 这里所讲的频数分布表是严格按照数值定义的精确列表, 而非传统意义上的按照一定的组距划分得出的简易频数表。

6.2.1 Frequencies过程的操作界面

执行【Analyze】/【Descriptive Statistics】/【Frequencies】命令, 弹出如图6-2所示对话框。

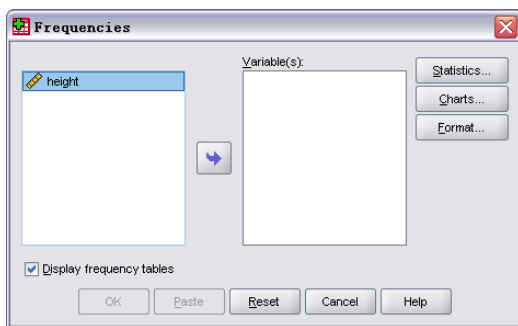


图 6-2 【Frequencies】对话框

该对话框主要用来定义频数表的特征。

1. 【Variables】

变量列表框。用于选择要产生频数表的变量, 可同时选择多个变量。此时, SPSS 就将分别产生多张频数表。

2. “Display frequency tables”

是否显示频数表, 系统默认为要显示。

3. 【Statistics】

单击该按钮, 弹出如图6-3所示的对话框。该对话框主要用于选择需要计算的统计量。Statistics 对话框将可选统计量分为4类, 分别为:

- ① 分位点 (Percentile Values)。
- ② 描述变量集中趋势的统计量 (Central Tendency)。
- ③ 描述变量离散程度的统计量 (Dispersion)。
- ④ 描述变量分布情况的统计量 (Distribution)。

另外,【Statistics】对话框上还有一个“Values are group midpoints”选项。该选项只有在选中“Percentile Values”组中选项时才有用,用来标识分位点是否恰好是变量的某个取值。

4. 【Charts】

单击图 6-2 中的【Charts】按钮,弹出如图 6-4 所示对话框。该对话框主要用于定义图形和图形中的数据。“Chart Type”组定义绘制的统计图类型,包括不绘制统计图、绘制条形图、饼图和直方图。如果选择绘制直方图的话,就会提示是否在直方图上添加正态曲线。“Chart Values”组用来选择条形图或饼图上的数据是显示变量值的频数还是百分比。

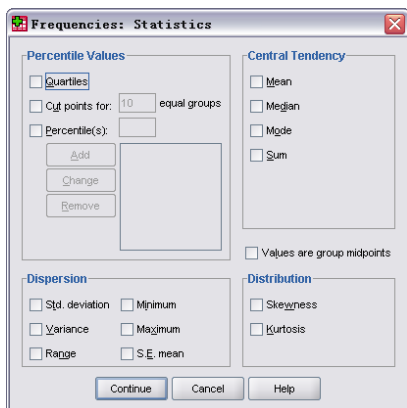


图 6-3 【Statistics】定义框

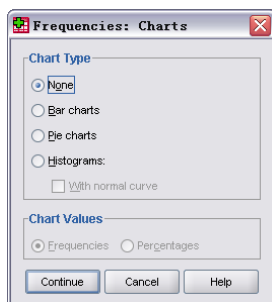


图 6-4 【Charts】对话框

5. 【Format】

单击图 6-2 上的【Format】按钮,弹出如图 6-5 所示对话框。该对话框主要用于定义频数表的输出。

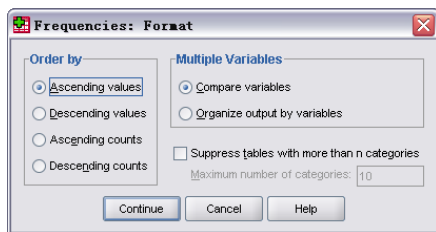


图 6-5 【Format】对话框

① Order by: 定义频数表中数据的显示方式。包括按照数值大小升/降序排列和按照频数大小升/降序排列 4 种情况。

② Multiple Variables: 定义当同时选择多个变量时,是否在同一个表中比较多个变量的统计量。选择“Compare variables”则要在同一个表中相互比较变量间的统计量。选择

“Organize output by variables” 则多个变量的统计量分别输出。

③ Suppress tables with more than n categories: 定义当频数表中个数大于 n 时不输出频数表，以避免输出过大的频数表。

6.2.2 引例

下面通过一个例子来介绍通过【Frequencies】过程来编制传统意义上的简易频数表的方法。

例 6.1 学生身高频数表。已知有某地 120 名 12 岁学生身高数据保存在数据文件“height.sav”中，编制其传统的简易频数表。

STEP 01 确定变量的最大值、最小值，以及全距。执行以下操作：

执行【Analyze】/【Descriptive Statistics】/【Frequencies】命令，弹出【Frequencies】对话框

【Variables】: height

单击【Statistics】按钮

弹出【Statistics】对话框

在【Statistics】对话框中选择要输出的统计量“Minimum”、“Maximum”、“Range”

在【Statistics】对话框中单击【Continue】

【Statistics】对话框定义完成

不选择“Display frequency tables”

定义不输出数值频数表

单击【OK】按钮

定义完成

执行以上操作之后，生成如表 6-4 所示结果。

表 6-4 身高的部分统计信息

Statistics		
height		
N	Valid	120
	Missing	0
Range		37.60
Minimum		122.70
Maximum		160.30

由表 6-4 可知，这批数据的最小值是 122.70，最大值是 160.30，全距为 37.6。拟选取 8 个实数点 126、130、134、138、142、146、150、154 将数据分成 9 组。

STEP 02 对数据进行分组。执行【Transform】/【Recode Into Different Variables】命令，按照 Step1 所确定的分组方式对变量进行分组，并且将生成的结果赋值给新变量“group”。具体的操作过程请参阅第 3 章。图 6-6 是执行以上操作之后的数据结构。

	height	group
1	128.10	2.00
2	134.10	4.00
3	126.00	1.00

图 6-6 分组后的学生身高数据

STEP 03 对新变量 “group” 绘制精确频数表。执行以下操作：

执行【Analyze】/【Descriptive Statistics】/【Frequencies】命令，弹出【Frequencies】对话框	
【Variables】：group	选择变量 “group”
单击【Charts】按钮	弹出【Charts】对话框
在【Charts】对话框中选择 “Histograms”	选择绘制直方图
选中 “With normal curve” 选项	在直方图上添加正态曲线
在【Charts】对话框单击【Continue】	【Charts】对话框定义完成
单击【OK】按钮	定义完成

生成的频数分布表如表 6-5 所示，生成的频数分布图如图 6-7 所示。

表 6-5 变量 group 的频数分布表

group					
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1.00	8	6.7	6.7	6.7
	2.00	6	5.0	5.0	11.7
	3.00	11	9.2	9.2	20.8
	4.00	20	16.7	16.7	37.5
	5.00	33	27.5	27.5	65.0
	6.00	22	18.3	18.3	83.3
	7.00	10	8.3	8.3	91.7
	8.00	5	4.2	4.2	95.8
	9.00	5	4.2	4.2	100.0
Total		120	100.0	100.0	

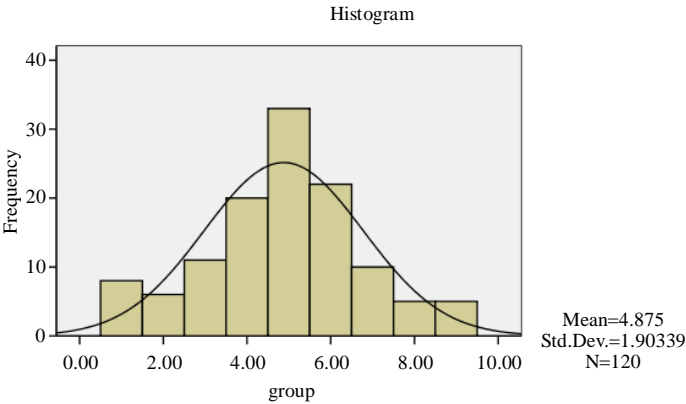


图 6-7 变量 group 的频数分布图

对于变量 group 而言，表 6-5 是精确数值的频数分布表。其中 Valid 表原始数据，Frequency 表对应数据的频数，Percent 为频数百分比，Valid Percent 为有效百分比，Cumulative Percent 为累计百分比。但是对于变量 “height” 而言，这就是它的传统意义上

按照分组方式确定的简易频数表了。从图形上看，学生身高的分布具有明显的正态性。

6.3 最基础的统计量分析——Descriptive过程

【Descriptive】过程主要用于输出变量的各类描述性统计量的值。通过上一节的学习可知，【Frequencies】过程同样可以做到这一点。在 SPSS 中，仍然将【Descriptive】过程单独列出来是为了方便只求描述性统计量而无需进行其他分析的问题。

6.3.1 Descriptive过程的操作界面

执行【Analyze】/【Descriptive Statistics】/【Descriptive】命令，弹出如图 6-8 所示对话框。

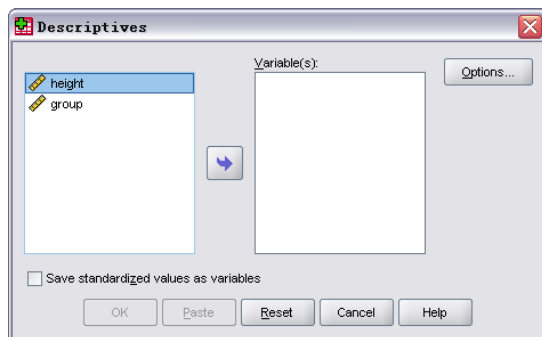


图 6-8 【Descriptives】对话框

【Descriptives】对话框十分简洁且看起来和【Frequencies】类似。该对话框主要由 3 部分组成。

1. 【Variables】

定义要分析的变量，可同时选择多个变量。

2. “Save standardized values as variables” 选项

定义是否将原始数据的 Z 变换（标准正态变换）结果存在数据文件中。Z 变换的公式为 $z_i = \frac{x_i - \bar{x}}{S}$ 。其中 \bar{x} 表变量的均值， S 表变量的标准差。如果选择该项，则数据文件中将自动生成一列名为“Z+原变量名”的新变量。

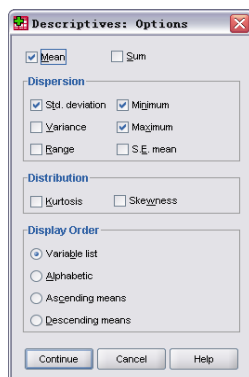


图 6-9 【Options】对话框

3. 【Options】

单击该按钮，弹出如图 6-9 所示的对话框。该对话框主要由 4 部分组成，用于选择要计算的统计量。

- ① 描述变量集中趋势的统计量（Mean&Sum）。
- ② 描述变量离散程度的统计量（Dispersion）。

③ 描述变量分布情况的统计量（Distribution）。

④ Display Order: 当【Variables】框选入多个变量时，选择变量分析结果的输出顺序。包括按数据文件中变量的排列顺序、按字母顺序、按均值升序/降序排列 4 种形式。

6.3.2 引例及结果解释

下面采用例 6.1 所用的数据文件“height.sav”进行统计分析。

例 6.2 学生身高的统计描述。

执行以下操作：

执行【Analyze】/【Descriptive Statistics】/【Descriptives】命令，弹出【Descriptives】对话框	
【Variables】: height	选择身高作为待分析变量
单击【Options】按钮	弹出【Options】对话框
在【Options】对话框中选择要输出的统计量	
在【Options】对话框单击【Continue】按钮	【Options】对话框定义完成
选中“Save standardized values as variables”选项	对变量进行 Z 变换
单击【OK】按钮	定义完成

执行以上操作之后，数据结构变成如图 6-10 所示。可以发现在数据文件中自动生成了一列新变量“Zheight”。该变量即是由变量“height”做标准正态变换得来的。

	height	Zheight	var
1	128.10	-1.50118	
2	134.10	-0.70025	
3	126.00	-1.78150	
4	133.40	-0.79369	

图 6-10 Z 变换之后的数据结构

表 6-6 所示为 Descriptives 过程在 SPSS Viewer 的输出结果。该结果包括变量值的个数、极值、均值、标准差、偏度和峰度信息。由于变量的偏度为 0.067，近似等于 0，因此可以认为其是对称的。峰度值也与 0 较为接近，说明变量的陡峭程度和正态分布差不多。

表 6-6 Descriptives 过程输出结果

Descriptive Statistics									
	N	Minimum	Maximum	Mean	Std.	Skewness		Kurtosis	
	Statistic	Statistic	Statistic	Statistic	Statistic	Statistic	Std.Error	Statistic	Std.Error
height	120	122.70	160.30	139.3458	7.49134	.067	.221	.135	.438
Valid N (listwise)	120								

6.4 探索性分析——Explore过程

与前面介绍的两个过程相比，【Explore】过程更加强大。它除了可以计算常见描述性统计量外，还可以给出一些简单的检验结果和图形，有助于读者进一步地分析数据。

6.4.1 Explore过程的操作界面

执行【Analyze】/【Descriptive Statistics】/【Explore】命令，弹出如图 6-11 所示对话框。

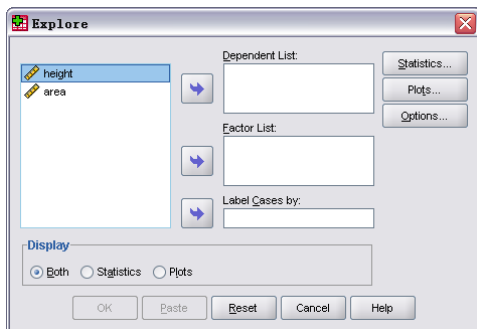


图 6-11 【Explore】对话框

该对话框主要由以下几部分组成。

- 【Dependent List】

选择待分析的变量，可以同时选择多个变量。

- 【Factor List】

选择分组变量，根据该变量取值不同，分组分析【Dependent List】框中的变量。可以不选，也可以选多个。

- 【Label Cases by】

选择标签变量，只能选一个。

- “Display” 单选框组

定义 SPSS View 窗口的输出结果。包括输出描述性统计结果、统计图或二者同时输出。

- 【Statistics】

单击该按钮，弹出如图 6-12 所示对话框。

该对话框主要用于选择要计算和输出的统计量。

① Descriptives: 计算一般的描述性统计量。在后面将通过例子给出具体包括哪些描述性统计量。

② M-estimators: 描述集中趋势的统计量。但是其受异常值的影响较小，因此在有异常值的情况下，可以用 M-estimators 的值代替均值。

③ Outliers: 分别输出 5 个极大值和极小值。

④ Percentiles: 输出变量 5%、10%、25%、50%、75%、90%、95% 分位数。

- 【Plots】

单击图 6-11 中的【Plots】按钮，弹出如图 6-13 所示的对话框。该对话框主要用于定义图形的特征。

① Boxplots: 定义箱图的输出。包括将变量按分组绘制、绘制到一起和不绘制 3 种情况。

② Descriptive: 定义是否输出茎叶图和直方图。

③ Normality plots with test: 选择是否进行正态检验，且是否输出相应的 Q-Q 图。

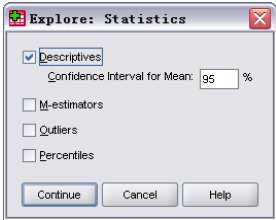


图 6-12 【Statistics】对话框

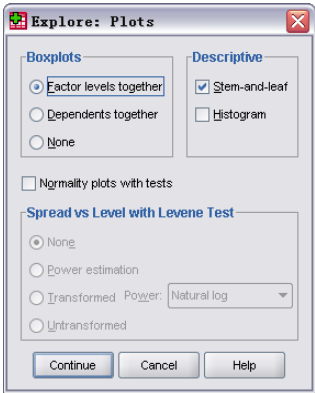


图 6-13 【Plots】对话框

④ Spread vs Level with Levene Test: 当选入分组变量时, 该功能组才有效。主要用于比较各组之间的离散程度是否一致, 其中 None 表示不进行比较; Power estimation 表示用于计算原始数据应该进行多少幂次方的计算才能使各组间的方差最齐; Transformed 提供了几种常见的数据转换方式; Untransformed 即不对数据进行转换。

• 【Options】

用于选择缺失值的处理方式。

6.4.2 引例及结果解释

下面通过一个例子来详细介绍【Explore】过程的输出结果。

例 6.3 学生身高的探索性分析。已知数据文件“height_1.sav”中有某地城市和农村 12 岁男生身高数据各 60 例, 利用【Explore】过程分组分析数据。

执行以下操作:

执行【Analyze】/【Descriptive Statistics】/【Explore】命令, 弹出【Explore】对话框	
【Dependent List】: height	选择身高作为待分析变量
【Factor List】: area	选择地区作为分组变量
【Statistics】:	对话框选中所有选项
【Plots】对话框:	
Boxplots : Factor levels together	选择将分组结果绘制在一张图上
Descriptive : Stem-and-Leaf	输出茎叶图
选中 Normality plots with test	进行正态性检验
Spread vs Level with Levene Test :	Power estimation 估计幂次使组间方差齐
单击【Continue】按钮	【Plots】对话框定义完成
单击【OK】按钮	定义完成

执行以上操作之后, 生成一大堆的统计图表, 按照 SPSS 的生成顺序一一利用这些图表进行探索性的统计分析。

表 6-7 显示了数据的基本情况。本例中每组有效数据各 60 例, 无缺失数据。

表 6-7 数据摘要

Case Processing Summary

area		Cases					
		Valid		Missing		Total	
		N	Percent	N	Percent	N	Percent
height	1.00	60	100.0%	0	.0%	60	100.0%
	2.00	60	100.0%	0	.0%	60	100.0%

表 6-8 输出的是描述性统计量。在本例中，由于利用变量“area”将数据分为两组，所以统计量的输出情况也分为两组。除了本章第一节介绍的统计量之外，表 6-8 还多生成了几个特殊的统计量，分别是均值的可信区间（Confidence interval for mean）、5%修正均数（Trimmed Mean）、四分位全距（Interquartile Range）。所谓的四分位全距即 3/4 分位点与 1/4 分位点之差。

表 6-8 描述性统计量

Descriptives

area			Statistic	Std.Error																																																																				
height	1.00	Mean	138.9917	1.03205																																																																				
		95% Confidence Lower Bound	136.9265																																																																					
		Interval for Mean Upper Bound	141.0568		5% Trimmed Mean	138.9519	Median	139.6500	Variance	63.908	Std. Deviation	7.99422	Minimum	123.10	Maximum	155.80	Range	32.70	Interquartile Range	9.83	Skewness	-.086	Kurtosis	-.441						2.00	Mean	139.7000	.90396	95% Confidence Lower Bound	137.8912	Interval for Mean Upper Bound	141.5088	5% Trimmed Mean	139.5204	Median	139.7000	Variance	49.029	Std. Deviation	7.00206	Minimum	122.70	Maximum	160.30	Range	37.60	Interquartile Range	8.15	Skewness	.344	Kurtosis	1.017															
		5% Trimmed Mean	138.9519		Median	139.6500	Variance	63.908	Std. Deviation	7.99422	Minimum	123.10	Maximum	155.80	Range	32.70	Interquartile Range	9.83	Skewness	-.086	Kurtosis	-.441						2.00			Mean	139.7000		.90396	95% Confidence Lower Bound	137.8912	Interval for Mean Upper Bound	141.5088	5% Trimmed Mean	139.5204	Median	139.7000	Variance	49.029	Std. Deviation	7.00206	Minimum	122.70	Maximum	160.30	Range	37.60	Interquartile Range	8.15	Skewness	.344	Kurtosis	1.017														
		Median	139.6500		Variance	63.908	Std. Deviation	7.99422	Minimum	123.10	Maximum	155.80	Range	32.70	Interquartile Range	9.83	Skewness	-.086	Kurtosis	-.441						2.00					Mean	139.7000			.90396	95% Confidence Lower Bound	137.8912	Interval for Mean Upper Bound	141.5088	5% Trimmed Mean	139.5204	Median	139.7000	Variance	49.029	Std. Deviation	7.00206	Minimum	122.70	Maximum	160.30	Range	37.60	Interquartile Range	8.15	Skewness	.344	Kurtosis	1.017													
		Variance	63.908		Std. Deviation	7.99422	Minimum	123.10	Maximum	155.80	Range	32.70	Interquartile Range	9.83	Skewness	-.086	Kurtosis	-.441						2.00							Mean	139.7000				.90396	95% Confidence Lower Bound	137.8912	Interval for Mean Upper Bound	141.5088	5% Trimmed Mean	139.5204	Median	139.7000	Variance	49.029	Std. Deviation	7.00206	Minimum	122.70	Maximum	160.30	Range	37.60	Interquartile Range	8.15	Skewness	.344	Kurtosis	1.017												
		Std. Deviation	7.99422		Minimum	123.10	Maximum	155.80	Range	32.70	Interquartile Range	9.83	Skewness	-.086	Kurtosis	-.441						2.00									Mean	139.7000					.90396	95% Confidence Lower Bound	137.8912	Interval for Mean Upper Bound	141.5088	5% Trimmed Mean	139.5204	Median	139.7000	Variance	49.029	Std. Deviation	7.00206	Minimum	122.70	Maximum	160.30	Range	37.60	Interquartile Range	8.15	Skewness	.344	Kurtosis	1.017											
		Minimum	123.10		Maximum	155.80	Range	32.70	Interquartile Range	9.83	Skewness	-.086	Kurtosis	-.441						2.00											Mean	139.7000						.90396	95% Confidence Lower Bound	137.8912	Interval for Mean Upper Bound	141.5088	5% Trimmed Mean	139.5204	Median	139.7000	Variance	49.029	Std. Deviation	7.00206	Minimum	122.70	Maximum	160.30	Range	37.60	Interquartile Range	8.15	Skewness	.344	Kurtosis	1.017										
		Maximum	155.80		Range	32.70	Interquartile Range	9.83	Skewness	-.086	Kurtosis	-.441						2.00													Mean	139.7000							.90396	95% Confidence Lower Bound	137.8912	Interval for Mean Upper Bound	141.5088	5% Trimmed Mean	139.5204	Median	139.7000	Variance	49.029	Std. Deviation	7.00206	Minimum	122.70	Maximum	160.30	Range	37.60	Interquartile Range	8.15	Skewness	.344	Kurtosis	1.017									
		Range	32.70		Interquartile Range	9.83	Skewness	-.086	Kurtosis	-.441						2.00															Mean	139.7000								.90396	95% Confidence Lower Bound	137.8912	Interval for Mean Upper Bound	141.5088	5% Trimmed Mean	139.5204	Median	139.7000	Variance	49.029	Std. Deviation	7.00206	Minimum	122.70	Maximum	160.30	Range	37.60	Interquartile Range	8.15	Skewness	.344	Kurtosis	1.017								
		Interquartile Range	9.83		Skewness	-.086	Kurtosis	-.441						2.00																	Mean	139.7000									.90396	95% Confidence Lower Bound	137.8912	Interval for Mean Upper Bound	141.5088	5% Trimmed Mean	139.5204	Median	139.7000	Variance	49.029	Std. Deviation	7.00206	Minimum	122.70	Maximum	160.30	Range	37.60	Interquartile Range	8.15	Skewness	.344	Kurtosis	1.017							
		Skewness	-.086		Kurtosis	-.441						2.00																			Mean	139.7000										.90396	95% Confidence Lower Bound	137.8912	Interval for Mean Upper Bound	141.5088	5% Trimmed Mean	139.5204	Median	139.7000	Variance	49.029	Std. Deviation	7.00206	Minimum	122.70	Maximum	160.30	Range	37.60	Interquartile Range	8.15	Skewness	.344	Kurtosis	1.017						
		Kurtosis	-.441							2.00																					Mean	139.7000											.90396	95% Confidence Lower Bound	137.8912	Interval for Mean Upper Bound	141.5088	5% Trimmed Mean	139.5204	Median	139.7000	Variance	49.029	Std. Deviation	7.00206	Minimum	122.70	Maximum	160.30	Range	37.60	Interquartile Range	8.15	Skewness	.344	Kurtosis	1.017					
								2.00																							Mean	139.7000												.90396	95% Confidence Lower Bound	137.8912	Interval for Mean Upper Bound	141.5088	5% Trimmed Mean	139.5204	Median	139.7000	Variance	49.029	Std. Deviation	7.00206	Minimum	122.70	Maximum	160.30	Range	37.60	Interquartile Range	8.15	Skewness	.344	Kurtosis	1.017				
						2.00																									Mean	139.7000													.90396	95% Confidence Lower Bound	137.8912	Interval for Mean Upper Bound	141.5088	5% Trimmed Mean	139.5204	Median	139.7000	Variance	49.029	Std. Deviation	7.00206	Minimum	122.70	Maximum	160.30	Range	37.60	Interquartile Range	8.15	Skewness	.344	Kurtosis	1.017			
	2.00	Mean	139.7000	.90396																																																																				
		95% Confidence Lower Bound	137.8912																																																																					
		Interval for Mean Upper Bound	141.5088																										5% Trimmed Mean	139.5204	Median	139.7000	Variance													49.029	Std. Deviation	7.00206	Minimum	122.70	Maximum	160.30	Range	37.60	Interquartile Range	8.15	Skewness	.344	Kurtosis	1.017												
		5% Trimmed Mean	139.5204																								Median	139.7000	Variance	49.029	Std. Deviation	7.00206	Minimum	122.70												Maximum	160.30	Range	37.60	Interquartile Range	8.15	Skewness	.344	Kurtosis	1.017																	
		Median	139.7000																						Variance	49.029	Std. Deviation	7.00206	Minimum	122.70	Maximum	160.30	Range	37.60	Interquartile Range											8.15	Skewness	.344	Kurtosis	1.017																						
		Variance	49.029																				Std. Deviation	7.00206	Minimum	122.70	Maximum	160.30	Range	37.60	Interquartile Range	8.15	Skewness	.344	Kurtosis	1.017																																				
		Std. Deviation	7.00206																		Minimum	122.70	Maximum	160.30	Range	37.60	Interquartile Range	8.15	Skewness	.344	Kurtosis	1.017																																								
		Minimum	122.70																Maximum	160.30	Range	37.60	Interquartile Range	8.15	Skewness	.344	Kurtosis	1.017																																												
		Maximum	160.30														Range	37.60	Interquartile Range	8.15	Skewness	.344	Kurtosis	1.017																																																
		Range	37.60												Interquartile Range	8.15	Skewness	.344	Kurtosis	1.017																																																				
		Interquartile Range	8.15										Skewness	.344	Kurtosis	1.017																																																								
		Skewness	.344								Kurtosis	1.017																																																												
		Kurtosis	1.017																																																																					

表 6-9 表示数据的 M 均值估计。在 SPSS 中，根据权系数的不同，共提供了 4 种估计方法。表 6-9 下方的注释分别给出了 4 种方法的权系数值。本例的数据也许不能看出 M 均值的稳健性。但是如果读者将本例某数据的小数点去掉，此时再比较均值和 M 均值的大小，发现对于有异常值的数据，M 均值估计法具有很好的稳定性。因此，如果由【Explore】过程计算出来的均值和 M 均值有较大的差距，那么用户就应当注意数据中是否有异常值了。

表 6-9 M 均值估计结果

M-Estimators				
area	Huber'S M-Estimator ^a	Tukey's Biweight ^b	Hampel's M-Estimator ^c	Andrews' Wave ^d
height 1.00	139.3861	139.4465	139.3063	139.4438
2.00	139.5890	139.4894	139.5227	139.4896

a. The Weighting constant is 1.339.

b. The weighting constant is 4.685.

c. The weighting constants are 1.700, 3.400, and 8.500

d. The Weighting constant is 1.340*pi.

表 6-10 给出了分位点信息。其中 Tukey's Hinges 表示的是绘制箱图时所用的分位点数据，它的计算方法和一般的百分位数略有不同。

表 6-10 分位点表

Percentiles									
area			Percentiles						
			5	10	25	50	75	90	95
Weighted	height	1.00	125.4100	126.1400	134.2250	139.6500	144.0500	150.3900	152.6700
Average (Definition 1)		2.00	128.2450	131.0800	135.4000	139.7000	143.5500	147.6800	154.1800
Tukey's Hinges	height	1.00			134.3500	139.6500	143.7000		
		2.00			135.5000	139.7000	143.4000		

表 6-11 给出了两组数据的极值信息，通过该表可以快速查找异常值。

表 6-11 极值表

Extreme Values					
area				Case Number	Value
height	1.00	Highest	1	55	155.80
			2	16	154.80
			3	10	152.70
			4	33	152.10
			5	56	150.70
		Lowest	1	37	123.10
			2	22	124.30
			3	42	125.40
			4	23	125.60
			5	57	126.00 ^a

续表

area		Case Number	Value
Highest	2.00	76	160.30
	1	115	156.90
	2	61	154.40
	3	77	150.00
	4	94	148.50
	5		
Lowest	1	63	122.70
	2	119	125.40
	3	89	128.20
	4	108	129.10
	5	62	130.30

a. Only a partial list of cases with the value 126.00 are shown in the table of lower extremes.

表 6-12 是正态性检验结果表。这里分别利用 Kolmogorov-Smirnov 检验和 Shapiro-Wilk 检验两种方法来确定变量是否服从正态分布。其中 Statistic 代表检验统计量的值，df 代表自由度，Sig. 代表显著水平。一般来说，Sig.>0.05 则代表接受假设。由于表中两种方法的 Sig.值均大于 0.05，因此接受变量服从正态分布的假设。

表 6-12 正态性检验表

Tests of Normality							
area		Kolmogorov-Smirnov ^a			Shapiro-Wilk		
		Statistic	df	Sig.	Statistic	df	Sig.
height	1.00	.080	60	.200*	.977	60	.328
	2.00	.076	60	.200*	.981	60	.490

*. This is a lower bound of the true significance.

a. Lilliefors Significance correction

表 6-13 是稳健的 Levene 方差齐次性检验结果。从上至下分别表示依赖于均值、中位数、中位数调整自由度，以及去掉极值的均值检验结果。此时由于 Sig.大于 0.05，可以认为方差是齐次的。

表 6-13 方差齐次性检验表

Test of Homogeneity of Variance					
		Levene Statistic	df1	df2	Sig.
height	Based on Mean	1.625	1	118	.205
	Based on Median	1.476	1	118	.227
	Based on Median and With adjusted df	1.476	1	117.530	.227
	Based on trimmed mean	1.622	1	118	.205

图 6-14 所示为地区 1 的身高茎叶图。茎叶图包括频数 (Frequency)、茎 (Stem) 和叶 (Leaf) 3 部分。下方的说明中给出了本图的茎宽为 10，每片叶子代表一个数据。对应到图形中的第一行，频数为 2，茎为 12，有两片叶子分别为 3 和 4。即指第一组里包括了两个数据，其近似值分别是 $12.3 \times 10 = 123$ 和 $12.4 \times 10 = 124$ 。同理，第二组包括 8 个数据，近似值分别为 125、125、126、126、127、127、127、128。

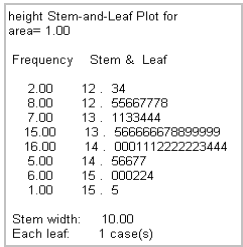


图 6-14 地区 1 的茎叶图

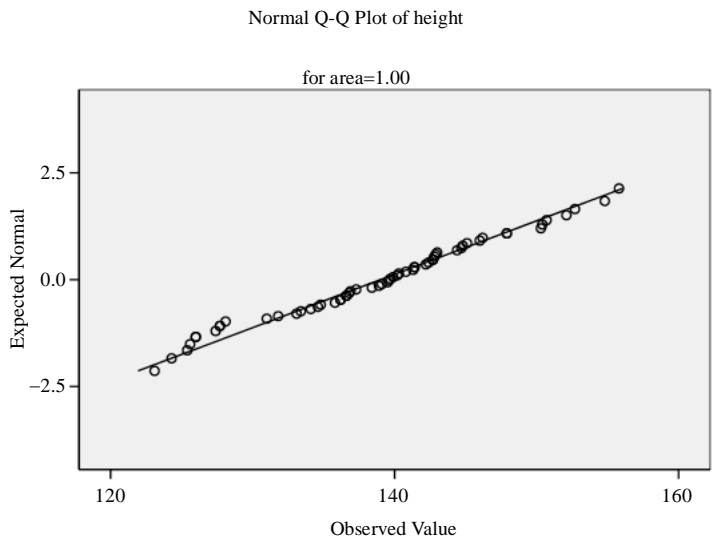


图 6-15 地区 1 身高数据的 Q-Q 图

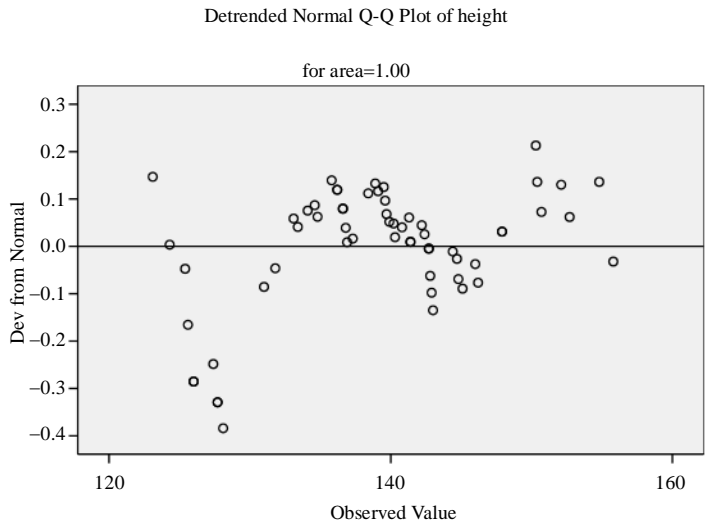


图 6-16 地区 1 身高数据的 Q-Q 去势图

图 6-17 所示为两个地区身高的箱图。由于前面【Plot】对话框中选择的是 Factor levels together，所以两地区的箱图绘制在一张图上。每一个箱体上方那条线的取值代表最大值，下方那条线的取值代表最小值。箱体自身的三条线从上到下分别代表 3/4 分位点、中位点、1/4 分位点的取值。地区 2 的箱图上有 3 组数据用“。”号标注出来，代表它们是离群值。“。”号旁的数据是指该离群值记录号。

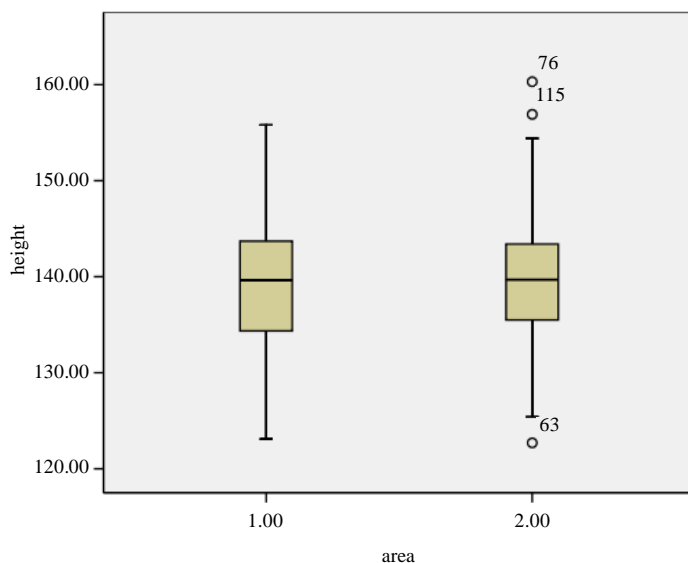
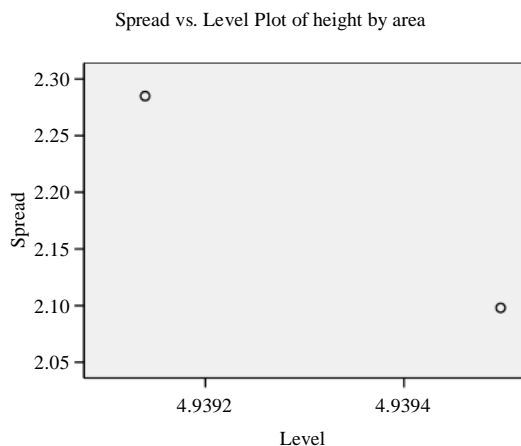


图 6-17 两地区身高箱图

通过图 6-18 来判断两组数据的离散程度是否一致。其中横轴为两地区身高中位数的自然对数，纵轴为两地区身高四分位数间距的自然对数。图形下方的注释给出了相应直线的斜率，以及使得两地区方差最齐的幂次转换估计值。



*Plot of LN of Spread vs LN of Level
Slope=-522.139 Power for transformation=523.139

图 6-18 散点图

6.5 列联表分析——Crosstabs过程

列联表给出了多个变量在不同取值下的数据分布，从而分析变量之间的相互关系。本节将通过例子详细介绍【Crosstabs】过程。

6.5.1 Crosstabs过程的操作界面

执行【Analyze】/【Descriptive Statistics】/【Crosstabs】命令，弹出如图 6-19 所示对话框。

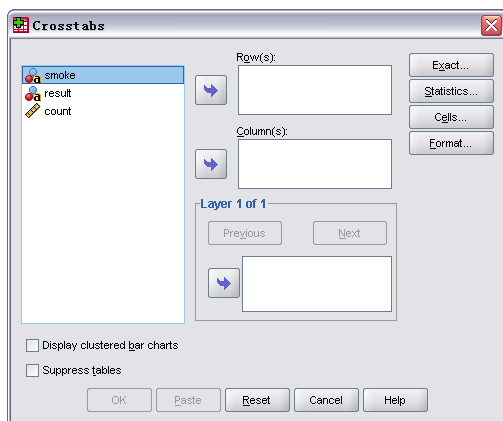


图 6-19 【Crosstabs】对话框

【Crosstabs】对话框主要由以下几部分组成。

- 【Rows】

选择列变量。

- 【Column】

选择行变量。

- 【Layer】

选择分层变量。用【Previous】和【Next】按钮控制分层的层数。

- 【Display clustered bar charts】

选择是否输出分组条图。

- 【Suppress tables】

选择是否禁止输出列联表。

- 【Exact】

单击【Exact】按钮，弹出如图 6-20 所示对话框。该对话框主要用于定义确切概率的计算。

① Asymptotic only: 只计算近似概率。

② Monte Carlo: 用 Monte Carlo 法计算精确概率，可自行设置置信度和抽样次数。

③ Exact: 在给定时间内计算精确概率的值, 如果超出给定时间则停止计算。

• 【Statistics】

单击图 6-19 中的【Statistics】按钮, 弹出如图 6-21 所示对话框。

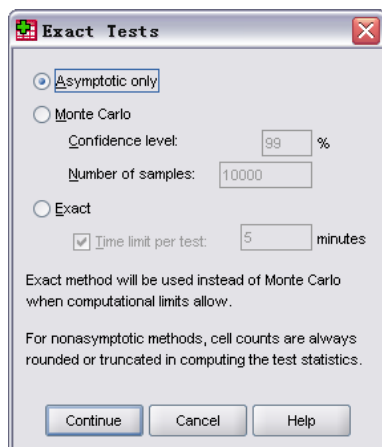


图 6-20 【Exact】对话框

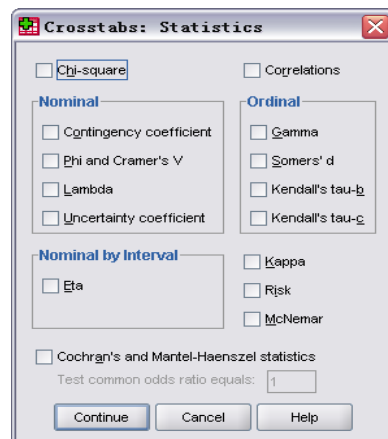


图 6-21 【Statistics】对话框

- ① Chi-square: 选择是否进行卡方检验。
- ② Correlations: 计算列联表的 Pearson 相关系数和 Spearson 相关系数。
- ③ Nominal: 定义分类变量的相关性指标, 包括如表 6-14 所示的 4 个指标。

表 6-14 Nominal 组的相关性指标

名 称	说 明	结果解释
Contingency coefficient	列联系数	取值在 (0, 1) 之间, 值越大说明两变量间的相关性越强
Phi and Cramer's V	Phi and Cramer's V 系数	在不同的卡方检验中, 取值范围不同。但是指标的绝对值越大, 说明变量间的相关性越强
Lambda	λ 系数	反映自变量对因变量的预测效果。取 1 时表示自变量可以很好地预测因变量, 取 0 时表示自变量和因变量之间完全没有可预测的关系
Uncertainty coefficient	不确定系数	以熵为标准反映一个变量对另一个变量的确定程度。取值为 1 时, 代表可由一个变量的信息完全确定另一个变量信息, 为 0 则代表两变量之间的信息完全没有关系

- ④ Ordinal: 定义有序变量的相关性指标, 包括如表 6-15 所示的 4 个指标。

表 6-15 Ordinal 组的相关性指标

名 称	结果解释
Gamma	取值在 (-1, 1) 之间, 取 1 或 -1 代表两变量完全一致或不一致, 取值为 0 代表两变量完全不相关
Somers' d	取值在 (-1, 1) 之间, 结果解释同上
Kendall's tau-b	取值在 (-1, 1) 之间, 结果解释同上
Kendall's tau-c	取值在 (-1, 1) 之间, 结果解释同上

⑤ Nominal by Interval: 用于分类变量的检验。其中 Eta 的平方表示因变量受不同因素影响的方差的比例。

⑥ **Kappa**: 内部一致性系数。通常 Kappa 大于 0.75 则认为两变量的一致性较好, 小于 0.4 则认为两变量的一致性较差。

⑦ **Risk**: 相对危险度, 主要用于医学统计上。

⑧ **McNemar**: 进行 McNemar 检验。该检验只有当行列数相等时才能用。

⑨ **Cochran's and Mantel-Haenszel statistics**: 独立性和齐次性检验。

• 【Cells】

单击图 6-19 中的【Cells】按钮, 弹出如图 6-22 所示的对话框。该对话框主要用于定义列联表中需要计算和输出的指标。

- **Count**: 定义输出频数, 包括实际观测数 (Observed) 和 (Expected) 理论数两项。
- **Percentages**: 定义需要计算的百分数, 包括行百分数 (Row)、列百分数 (Column) 和总体百分数 (Total)。
- **Residuals**: 定义输出的残差。包括非标准化残差, 即实际数—理论数之差 (Unstandardized)、标准化残差, 即 (实际数—理论数之差)/实际数 (Standardized) 以及调节标准化残差 (Adjusted Standardized)。
- **Noninteger Weights**: 当频数因为加权而变成小数的时候, 选择该项对频数进行取整。主要包括了如表 6-16 所示的 5 种取整方法。

表 6-16 Noninteger Weights 取整方法

名 称	方 法
Round cell counts	对频数进行四舍五入取整
Round case weights	对加权样本在使用前进行四舍五入取整
Turncate cell counts	对频数进行舍位取整
Turncate case weights	对加权样本在使用前进行舍位取整
No adjustments	不调整

• 【Format】

单击图 6-19 中的【Format】按钮, 弹出如图 6-23 所示的对话框。该对话框主要用于定义行变量的排列方式。

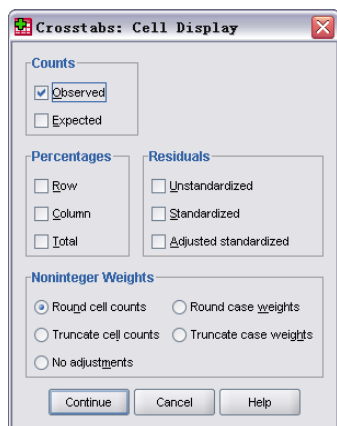


图 6-22 【Cells】对话框

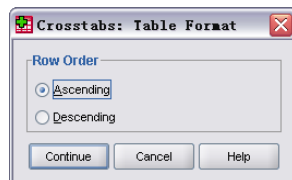


图 6-23 【Format】对话框

6.5.2 引例

四格表卡方检验和 $R \times C$ 卡方检验是【Crosstabs】过程中最常用的功能。四格表卡方检验即针对 2×2 的列联表。而 $R \times C$ 卡方检验则是针对 $m \times n$ 的列联表（ m 和 n 中至少有一个大于 2）。在 SPSS 中，四格表卡方检验和 $R \times C$ 卡方检验在操作上是类似的。下面通过一个例子来详细介绍四格表卡方检验。

例 6.4 吸烟习惯与患病率的关系。调查 339 名 50 岁以上吸烟习惯与患慢性气管炎病的关系，如表 6-17 所示。试问吸烟者与不吸烟者的慢性气管炎患病率是否有所不同。（数据来源：《概率论》，复旦大学出版社）

表 6-17 吸烟与气管炎关系的调查数据表

	患慢性气管炎者	未患慢性气管炎者	合 计	患病率
吸烟	43	162	205	21.0
不吸烟	13	121	134	9.7
合计	56	283	339	16.5

STEP 01 建立如图 6-24 所示数据文件“smoke.sav”。

	smoke	result	count
1	是	患病	43.00
2	是	健康	162.00
3	否	患病	13.00
4	否	健康	121.00

图 6-24 调查数据表的 SPSS 数据录入情况

STEP 02 对数据进行预处理。执行【Data】/【Weight Cases】命令，弹出如图 6-25 所示的【Weight Cases】对话框。选中“Weight cases by”单选框，将变量“count”放入【Frequency Variable】框中，单击【OK】按钮。完成对数据的预处理，将“count”设为频数变量。

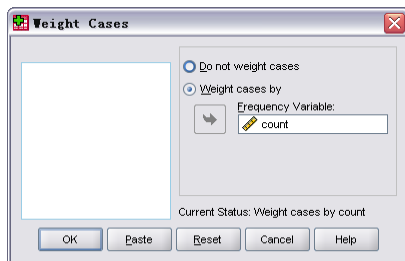


图 6-25 【Weight Cases】对话框

STEP 03 列联表分析。

执行【Analyze】/【Descriptive Statistics】/【Crosstabs】命令，弹出【Crosstabs】对话框	
【Rows】: smoke	选择是否吸烟作为行变量
【Columns】: result	选择是否患病作为列变量
选中【Display clustered bar charts】命令	绘制分组条图

【Statistics】对话框：

选择“Chi-square”

选择进行卡方检验

单击【Continue】按钮

【Statistics】对话框定义完成

单击【OK】按钮

定义完成

6.5.3 结果解释

例 6.4 在执行 6.5.2 的操作之后生成表 6-18～表 6-20，下面对这些表中的数据信息进行解释。
表 6-18 给出了数据基本信息。

表 6-18 数据摘要

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
smoke*result	339	100.0%	0	.0%	339	100.0%

表 6-19 给出了数据的 2×2 列联表。这与原始数据在形式上是基本一致的。

表 6-19 列联表

smoke*result Crosstabulation				
Count		result		Total
		患 病	健 康	
smoke	否	13	121	134
	是	43	162	205
Total		56	283	339

表 6-20 是数据的卡方检验结果，共使用了 5 种检验方法，依次为 Pearson 卡方检验（Pearson Chi-Square）、连续性校正卡方检验（Continuity Correction）、似然比卡方检验（Likelihood Ratio）、Fisher's 精确概率检验（Fisher's Exact Test）、有效记录数检验（N of Valid Cases）。计算的统计量主要包括检验统计量（Value）、自由度（df）、双测近似概率（Asymp.Sig.-2-Sided）、双测精确概率（Exact Sig.-2-Side）和单测精确概率（Exact.Sig.-1-Sided）。

表 6-20 卡方检验结果

Chi-Square Tests					
	Value	df	Asymp. Sig. (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	7.469 ^b	1	.006		
Continuity Correction ^a	6.674	1	.010		
Likelihood Ratio	7.925	1	.005		
Fisher's Exact Test				.007	.004
N of Valid Cases	339				

a. Computed only for a 2×2 table

b. 0 cells (.0%) have expected count less than 5. The minimum expected count is 22.14.

从表 6-20 可以看出，各种检验方法显著水平都是远远小于 0.05 的。可见，拒绝吸烟和患病是独立的假设，即认为吸烟和患支气管炎是相关的。

表 6-20 下方注释 b 为“0 单元格的期望值小于 5，表格的最小期望值为 22.14”。注释 b 主要用于决定选择何种卡方检验方法的结果。

注意 本例各类卡方检验的结果是一致的，所以避免了选择何种检验方法这个问题。但是在实际问题中，对于检验方法的选择是不可能回避的问题。一般而言，有如下准则：

- $n \geq 40$ 且 $T \geq 5$ ，用 Pearson 卡方检验
- $n \geq 40$ 且 $1 \leq T < 5$ ，用连续性校正卡方检验
- $n < 40$ 或 $T < 1$ ，用 Fisher's 精确概率检验

其中 n 为样本数， T 为期望值。

图 6-26 相当于是表 6-19 的直观表示。本例虽然不能直接根据图形下结论。但是对于某些问题，图形就能很明显地说明问题了。

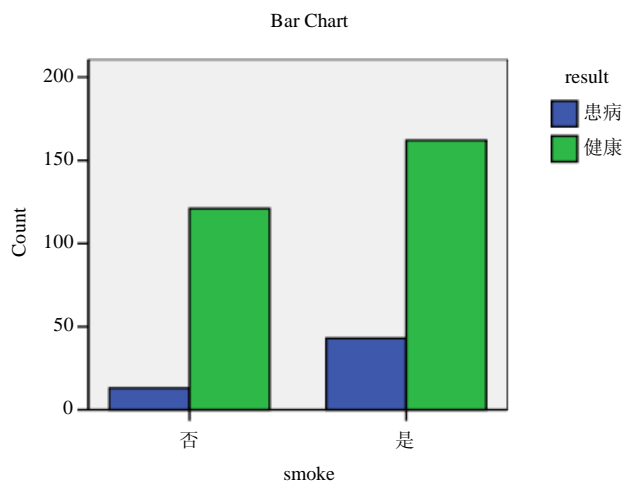


图 6-26 各组状况的分组条图

6.6 相对比描述——Ratio过程

在实际问题中，研究者有时希望除了了解变量自身的统计特征外，还希望得到两个变量相对比之间的统计描述。当然，这可以通过【Transform】菜单下的【Compute Variable】过程对两个变量做除法形成一个新变量，然后分析新变量的统计特征来得到。但是，在 SPSS 中可以直接通过【Ratio】过程来分析两个变量之间的相对比关系，并且可以得到多于第一种方法的信息。

6.6.1 Ratio过程的操作界面

执行【Analyze】/【Descriptive Statistics】/【Ratio】命令，弹出如图 6-27 所示对话框。

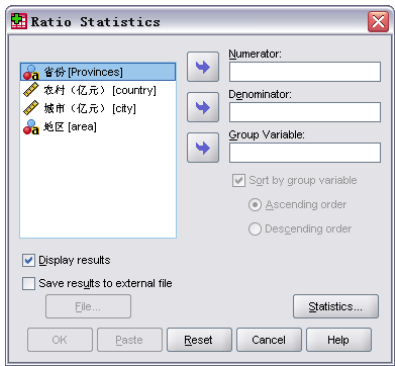


图 6-27 【Ratio】对话框

该对话框主要由以下几部分组成。

- **【Numerator】**
选择相对比中作分子的变量。
- **【Denominator】**
选择相对比中作分母的变量。
- **【Group Variable】**
选择分组变量。
- “Sort by group variable” 单选框组
当有分组变量时，选择分组变量的排序方式。
- **【Display result】**
选择是否显示结果。
- **【Save results to external file】**
选择是否将分析结果存入外部文件。
- **【Statistics】**
单击**【Statistics】**按钮，弹出如图 6-28 所示对话框。

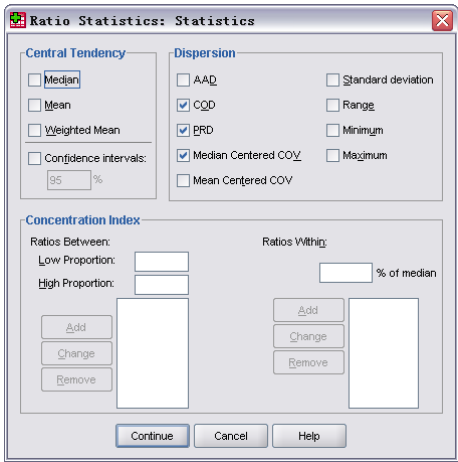


图 6-28 【Statistics】对话框

该对话框主要由以下几部分组成。

① **Central Tendency**: 描述相对比集中趋势的统计量, 包括中位数 (Median)、均值 (Mean)、加权均值 (Weighted mean), 并且可以通过下方的 “Confidence intervals” 设定输出相应指标的置信区间。

② **Dispersion**: 描述相对比离散程度的统计量。除了标准差、全距和极值之外, 还包括如表 6-21 所示的相对比的特定统计指标。

表 6-21 Dispersion 统计指标

SPSS 中表示	名 称	计算公式
ADD	平均绝对离差	$ADD = \frac{\sum_{i=1}^n r_i - r_{1/2} }{n}$
COD	离散系数	$COD = \frac{ADD}{r_{1/2}}$
PRD	价格相对微分	$PRD = \frac{\bar{r}}{r_w}$
Median centered COV	基于中位数的变异系数	$\sqrt{\frac{1}{n-1} \sum_{i=1}^n (r_i - r_{1/2})^2}$
Mean centered COV	基于均值的变异系数	$\sqrt{\frac{1}{n-1} \sum_{i=1}^n (r_i - \bar{r})^2}$

表格说明: r_i 表示第 i 个相对比, \bar{r} 表示相对比的均值, r_w 表示相对比的加权均值, $r_{1/2}$ 表示相对比的中位数。

③ **Concentration index**: 描述相对比在某一区间类的比例。其中 “Ratios Between” 是求相对比取值在上下界之间的数目占总数的比例。“Ratios of median” 是求相对比在中位数的百分比之内的数目占总数的比例。

6.6.2 引例及结果解释

例 6.5 各地区城乡居民消费水平比较。

已知有 2005 年各省城乡居民消费水平数据保存在数据文件 “xiaofei.sav” 内, 其数据结构如图 6-29 所示。试按地区对各省城乡消费水平之比进行分析, 并比较不同地区之间城乡消费水平是否有较大差异。(数据来源:《2005 中国统计年鉴》, 中国统计出版社)

	Provinces	country	city	area
1	北 京	172.15	1182.08	华北
2	天 津	177.46	628.65	华北
3	河 北	1368.78	1250.40	华北

图 6-29 城乡消费水平

执行以下操作:

执行【Analyze】/【Descriptive Statistics】/【Ratio】命令, 弹出【Ratio】对话框
【Numerator】: city 选择 “city” 作为相对比中的分子变量

【Denominator】：country	选择“country”作为相对比中的分母变量
【Group Variable】：area	按地区分组计算相对比
【Statistics】对话框：	
选中“Median”、“Mean”、“AAD”、“COD”、“PRD”、“Median centered COV”	
单击【Continue】按钮	统计量定义完成
单击【OK】按钮	定义完成

执行以上操作之后，生成数据摘要如表 6-22 所示。

表 6-22 所示是数据的基本信息。从表中可以看出各组有多少数据。在本例中，所有数据都纳入了分析。

表 6-22 数据摘要

Case Processing Summary			
		Count	Percent
地区	东北	3	9.7%
	华北	5	16.1%
	华南	6	19.4%
	华中	7	22.6%
	西北	5	16.1%
	西南	5	16.1%
Overall		31	100.0%
Excluded		0	
Total		31	

表 6-23 给出了城市农村消费水平相对比的一些基本统计量。

表 6-23 城市农村消费水平相对比

Ratio Statistics for 城市（亿元）/农村（亿元）							
Group	Mean	Median	AAD（Average Absolute Deviation）	Price Related Differential	COD （Coefficient of Dispersion）	Coefficient of Variation	
						Mean Centered	Median Centered
东北	3.602	3.486	.144	1.000	.041	6.6%	7.9%
华北	3.165	2.283	1.455	1.655	.637	71.7%	108.3%
华南	1.610	1.107	.709	.910	.640	61.4%	102.2%
华中	2.24g	.998	1.423	1.603	1.426	131.2%	325.4%
西北	1.586	1.596	.186	1.044	.116	16.5%	16.4%
西南	1.270	1.449	.373	1.064	.257	37.8%	35.9%
Overall	2.139	1.596	1.110	1.290	.696	84.4%	118.3%

分析表 6-23 中数据可以发现，东北地区相对比的均值和中位数都很大，说明该地区的城乡消费水平差距较大。东北和西北地区的 AAD 和 COD 较小，相对比离散程度低。华北和华中地区的 AAD 和 COD 较大，相对比离散程度高。说明东北地区 and 西北地区各省内部的城乡消费水平较一致，而华北地区 and 华中地区内部各省的城乡消费水平差距较大。这一

点也可以从实际情况中得到合理的解释。以东北地区和华北地区为例，东三省的经济水平较一致，所以各省城乡消费水平的比例也比较一致，其相对比离散度低。而对于华北地区，显然北京和山西、内蒙古等省的经济情况有较大差距，所以反映到城乡消费水平的比例上看也会有较大的离散程度。这说明我们用【Ratio】过程得出的结论是和实际情况相吻合的。

6.7 本章小结

本章介绍了 SPSS 描述性统计分析【Descriptive Statistics】子菜单，详细介绍了以下几个过程：

- Frequencies 过程，频数分布表分析；
- Descriptive 过程，最基础的统计量分析；
- Explore 过程，探索性分析；
- Crosstabs 过程，列联表分析；
- Ratio 过程，相对比描述。

本章的内容都相对比较简单，通过本章的学习，读者可以掌握如何利用 SPSS 软件进行最基础的描述性统计分析。

第 7 章 均值比较与t检验

t 检验是最常用的假设检验方法之一。SPSS 的【Compare Means】子菜单包括了各类 t 检验方法。本章将通过实际例子学习 t 检验的一般步骤及其在 SPSS 中的实现。本章内容包括：

- t 检验简介
- 均值描述——Means 过程
- 单样本 t 检验——One-Sample T Test 过程
- 独立两样本 t 检验——Independent-Sample T Test 过程
- 配对样本 t 检验——Paired-Sample T Test 过程


7.1 t检验简介

本节将概括性地介绍 t 检验的基本概念、一般步骤和类型。对于各类具体的 t 检验方法，将在本章的后续几节给出详细的介绍。

7.1.1 t检验的概念及一般步骤

简而言之，t 检验就是检验统计量为 t 的假设检验，主要用来检验均值之间的关系。所谓假设检验即是在统计推断中，根据样本观测值，检验总体参数或分布的假设是否正确的一种统计学方法。作为一类特殊的假设检验，t 检验的一般步骤与假设检验是一致的。通常包括以下几步。

- STEP 01** 根据实际问题提出原假设 H_0 与备假设 H_1 。
- STEP 02** 选择统计量 t 作为检验统计量，并在 H_0 成立的条件下确定 t 的分布。
- STEP 03** 选择显著性水平 α ，并根据统计量 t 的分布表查确定临界值及 H_0 的拒绝域。
- STEP 04** 根据样本值计算统计量的值，并将其与临界值做比较。
- STEP 05** 下结论：若统计量的值落入拒绝域内，则拒绝 H_0 ；否则，不拒绝 H_0 。

 **注意** 显著性水平的选取十分重要。对于同一个假设检验，选取不同，则可能会得出完全不同的结论。在实际问题中，通常取 0.05 或 0.01。

7.1.2 t检验的类型

在统计学中，通常将 t 检验分为 4 类：样本均值与总体均值比较的 t 检验、独立两样

本均值比较的 t 检验、配对设计的差数均值与总体均值 0 的 t 检验，以及独立两样本几何均值比较的 t 检验。在 SPSS 中，这几类检验方法都可以由如图 7-1 所示的【Compare Means】子菜单实现。

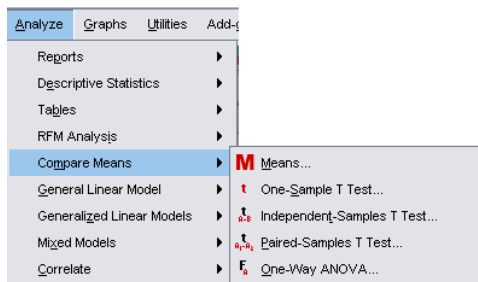


图 7-1 【Compare Means】子菜单

【Compare Means】子菜单主要通过均值比较，实现 t 检验和单因素方差分析，各过程的主要作用如下。

- ① Means：分组计算样本的描述性统计量。
- ② One-Sample T Test：单样本 t 检验，即比较样本均值和总体均值的 t 检验。
- ③ Independent-Sample T Test：独立两样本 t 检验，即比较两独立样本均值的 t 检验。
- ④ Paired-Sample T Test：配对样本 t 检验，即比较配对设计的差数均值与 0 的 t 检验。
- ⑤ One-Way ANOVA：单因素方差分析，这一过程将放在下一章方差分析中介绍。

在本章将通过具体的例子来学习前 4 类过程。

注意 U 检验也是一种常用的假设检验方法。当样本的个数较多时， t 检验计算较为烦琐，此时通常采用 U 检验。但是对计算机而言，大样本 t 检验的计算也是轻而易举的，所以仍然可以采用 t 检验。这也是在 SPSS 中没有专门集成 U 检验模块的原因。

7.2 均值描述——Means过程

与上一章类似，【Compare Means】子菜单中的【Means】过程也是用来计算各类描述性统计量的。但是在计算的过程中，【Means】过程能够按照给定变量分组计算并输出最后结果。同时【Means】过程能够直接输出方差分析结果，这显然比上一章的各个过程的功能更加强大。

7.2.1 Means过程的操作界面

执行【Analyze】/【Compare Means】/【Means】命令，弹出如图 7-2 所示对话框。该对话框主要用来选择待分析变量及其分组变量，各项的具体功能如下。

- 【Dependent List】

选择待分析变量。

- “Layer” 组

定义分组变量，主要包括以下3个按钮。

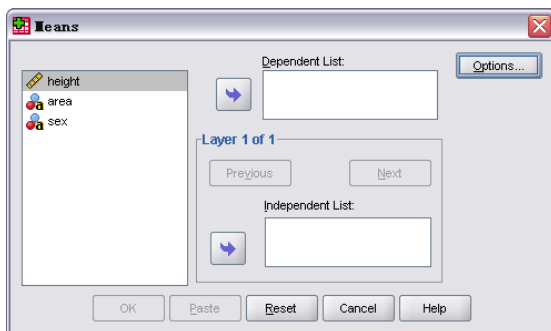


图 7-2 【Means】对话框

① **【Independent List】**: 选择分组变量。可定义多层分组变量，每层分组变量中也可以有多个变量。

② **【Previous】**: 选择上一层的分组变量。

③ **【Next】**: 选择下一层的分组变量。

- **【Options】**

单击该按钮，弹出如图 7-3 所示对话框。该对话框主要由以下 3 部分组成。

① **Statistics 框**: 列出可以选择的描述性统计量。由于这些统计量的具体含义在上一章已经详细介绍了，所以此处不再赘述。

② **Cell Statistics 框**: 选择要输出的统计量。默认输出均值 (Mean)、样本个数 (Number of Cases) 和标准差 (Standard Deviation)。

③ **Statistics for First Layer**: 定义是否进行分组第一层变量的方差分析 (Anova table and eta) 和线性检验 (Test for linearity)。

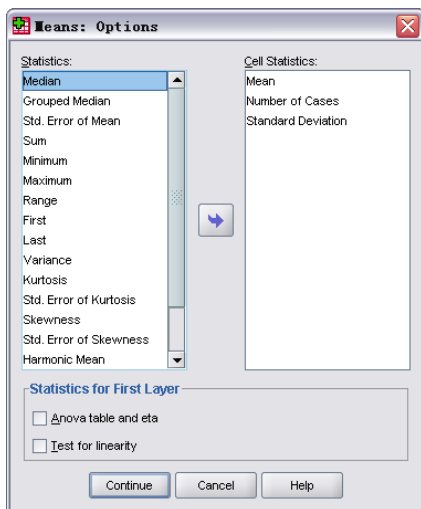


图 7-3 【Options】对话框

7.2.2 引例及结果解释

下面通过一个例子来介绍【Means】过程的操作及其结果。

例 7.1 已知从甲、乙两地各抽取 60 名 12 岁的学生，其中男女各占一半，其身高数据保存在数据文件“height.sav”中，利用【Means】过程比较身高是否受地区和性别的影响。

执行以下操作：

执行【Analyze】/【Compare Means】/【Means】命令，弹出【Means】对话框	
【Dependent List】：height	选择分析身高
【Independent List】：sex、area	性别和地区作为同层分组变量
单击【Options】按钮	弹出【Options】对话框
在【Options】对话框中选中 Anova table and eta，其余采用默认值	选择进行方差分析
在【Options】对话框单击【Continue】按钮	【Options】对话框定义完成
单击【OK】按钮	定义完成

执行以上操作之后，分别生成表 7-1～表 7-4 所示信息数据。

表 7-1 是数据的基本信息。从表格中可以看出 120 组数据全部有效。

表 7-1 数据摘要

Case Processing Summary

	Cases					
	Included		Excluded		Total	
	N	Percent	N	Percent	N	Percent
Height*sex*area	120	100.0%	0	.0%	120	100.0%

表 7-2 显示的是按性别分组的身高基本信息。由于首先选择的分组变量是“sex”，所以按性别分组的信息出现在按地区分组的信息的前面。从表 7-2 可以发现男女学生身高的均值和标准差都比较接近。

表 7-2 按性别分组的基本信息

Report

height			
Sex	Mean	N	Std.Deviation
男	139.8150	60	7.48243
女	138.8100	60	7.41073
Total	139.3125	120	7.43246

表 7-3 是性别的单因素方差分析表。在下一章将会详细介绍方差分析。表中的显著性水平 Sig.值远大于 0.05，说明男女学生身高没有显著性差异。

表 7-3 单因素方差分析表

ANOVA Table						
		Sum of Squares	df	Mean Square	F	Sig.
Height*sex	Between Groups (Combined)	30.301	1	30.301	.546	.461
	Within Groups	6543.431	118	55.453		
	Total	6573.731	119			

表 7-4 是身高与性别的相关性度量表。此时的 Eta 和 Eta Squared 取值都很小。说明性别和身高的相关性很小。这也和单因素方差分析表的结论是一致的。

表 7-4 身高与性别的相关性度量表

Measures of Association		
	Eta	Eta Squared
height*sex	.068	.005

表 7-2~表 7-4 是按照性别分组得出的分析结果。在本例中，由于分层变量同时选择了“sex”和“area”，所以还会输出按照地区分类的 3 张表。表格同表 7-2~表 7-4 完全类似，所以此处就不再一一解释了。

7.2.3 分组变量的层次说明

【Means】对话框上“Layer”组的分组变量的层次的定义方法是容易困扰初学者的问题之一。选择不同层次的分组变量，在输出结果上到底有何差别？这里不妨仍然用例 7.1 中的数据来做一个小小的变换。执行以下操作：

执行【Analyze】/【Compare Means】/【Means】命令，弹出【Means】对话框	
【Dependent List】: height	选择分析身高
【Independent List】: sex	选择性别作为第一层分组变量
单击【Next】按钮	准备定义下一层分组变量
【Independent List】: area	选择地区作为第二层分组变量
单击【OK】按钮	定义完成

执行以上操作之后生成的基本信息如表 7-5 所示。

表 7-5 两层分组变量下的基本信息

Report				
height				
sex	area	Mean	N	Std. Deviation
男	甲	139.8700	30	8.36372
	乙	139.7600	30	6.62917
	Total	139.8150	60	7.48243

续表

sex	area	Mean	N	Std. Deviation
女	甲	138.1133	30	7.64662
	乙	139.5067	30	7.22887
	Total	138.8100	60	7.41073
Total	甲	138.9917	60	7.99422
	乙	139.6333	60	6.87766
	Total	139.3125	120	7.43246

在例 7.1 中，由于将两个分组变量“sex”和“area”定义在同一层内，即二者是平等的关系，所以会分别按照性别和地区分组输出两张基本信息表。而在这里，由于将“sex”作为第一层分组变量，“area”作为第二层分组变量，二者之间是有层次关系的，所以最后输出的是先按性别分组，在同一性别内再按地区分组的一张基本信息表。

需要注意的是，此时虽然有两层分组变量。但是在做方差分析的时候，仅对第一层变量即“sex”作单因素方差分析。

7.3 单样本t检验——One-Sample T Test过程

单样本 t 检验是比较样本均值和总体均值的 t 检验。本节通过例子介绍单样本 t 检验的一般步骤、操作界面及结果解释。

7.3.1 单样本t检验的一般步骤

同一般的 t 检验一样，单样本 t 检验主要包括以下几步。

STEP 01 提出原假设 H_0 与备选择假设 H_1 。在单样本 t 检验中， H_0 为 $\mu = \mu_0$ ，即假设样本均值与总体均值的差异是由抽样误差造成的。 H_1 为 $\mu \neq \mu_0$ ，即假设样本均值同总体均值存在除了抽样误差之外的其他误差。

STEP 02 选取统计量 t 作为检验统计量，在 H_0 成立的条件下其应服从自由度为 $n-1$ 的 t 分布。其中 n 为样本个数。

STEP 03 选择显著性水平 α 。 α 是指若假设 H_0 为真，是被拒绝的概率。并根据统计量 t 的分布查表确定临界值 t_0 及 H_0 的拒绝域。

STEP 04 求样本均值 \bar{X} 、标准差 S ，计算检验统计量。此时有

$$t = \frac{\bar{X} - \mu_0}{S / \sqrt{n}} \quad (7.1)$$

STEP 05 下结论：若检验统计量 $|t| > t_0$ 就拒绝 H_0 ；否则，不拒绝 H_0 。

注意 (1) 检验统计量未落入拒绝域内，仅仅是不拒绝它，并不代表就一定要接受它。

(2) SPSS 中的 Sig. 值是指我们通常所说的 P 值。在 SPSS 中，通常若 $\text{Sig.} > 0.05$ ，则接受原假设，否则不接受。

(3) 样本来自的总体要服从正态分布。

7.3.2 One-Sample T Test过程的操作界面

执行【Analyze】/【Compare Means】/【One-Sample T Test】命令，弹出如图 7-4 所示对话框。

【One-Sample T Test】对话框十分简洁，主要由以下几部分组成。

- 候选变量框

即左侧变量列表框，该变量框只显示可以进行 t 检验的变量。

- 【Test Variables】

选择进行 t 检验的变量，可以同时选入多个变量。

- 【Test Value】

输入总体均值，即前面提到的 μ_0 。

- 【Options】

单击【Options】按钮，弹出如图 7-5 所示的对话框。

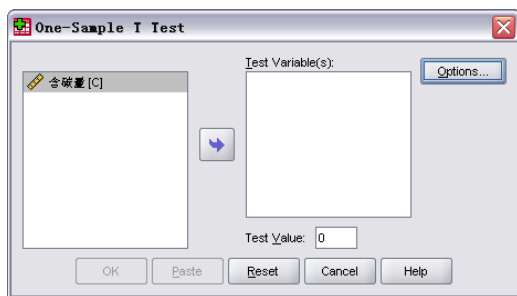


图 7-4 【One-Sample T Test】对话框

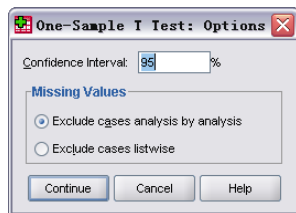


图 7-5 【Options】对话框

该对话框主要包含如下选项。

① Confidence Interval: 设置样本均值与总体均值之差的置信区间，默认为 95%。

② Missing Values: 设置缺失值处理方式。有两种处理方法：一种方法是仅当数据要分析的变量值缺失时才剔除该数据 (Exclude cases analysis by analysis)；另一种方法是只要数据中有变量值缺失就剔除该数据 (Exclude cases listwise)。为了保留更多的数据样本，默认为前者。

7.3.3 引例及结果解释

通过一个例子来介绍【One-Sample T Test】过程的操作及其结果。

例 7.2 已知某炼铁厂铁水含量符合均值为 4.53 的正态分布，某日随机测定了 9 炉铁水，含碳量如表 7-6 所示。

表 7-6 铁水含碳量抽样数据

抽 样	1	2	3	4	5	6	7	8	9
含 碳 量	4.43	4.50	4.58	4.42	4.47	4.60	4.53	4.46	4.42

问该日铁水平均含碳量是否仍为 4.53。

这是一个典型的比较样本均值和总体均值的 t 检验问题。执行以下操作。

STEP 01 建立数据文件“Fe.sav”。用变量“C”表示含碳量，录入数据。

STEP 02 进行单样本 t 检验，执行以下操作：

执行【Analyze】/【Compare Means】/【One-Sample T Test】命令，弹出【One-Sample T Test】对话框

【Test Variables】：C 设定含碳量 C 为待检验变量

【Test Value】：4.53 设定总体均值为 4.53

单击【OK】按钮 定义完成

执行以上操作之后，将生成表 7-7 和表 7-8。

表 7-7 给出了数据的基本的描述性统计量。包括样本数（N）、均值（Mean）、标准差（Std.Deviation）、标准误均值（Std.Error Mean）。在本例中，样本均值为 4.49，与总体均值 4.53 还是比较接近的。

表 7-7 单样本统计表

One-Sample Statistics				
	N	Mean	Std. Deviation	Std. Error Mean
含碳量	9	4.4900	.06764	.02255

表 7-8 是单样本 t 检验表。包括总体均值（Test Value）、检验统计量（ t ）、自由度（ df ）、双侧检验的显著性水平（Sig. (2-tailed)）、样本均值与总体均值之差（Mean Difference）和均值差的置信区间（Confidence Interval of the Difference）。在本例中， t 检验的统计量取值为 -1.774。由于双侧 t 检验的显著性水平 Sig. 取值为 0.114，大于 0.05，所以不拒绝假设 H_0 ，即认为样本均值与总体均值之差可能是由抽样误差所造成的。

表 7-8 单样本 t 检验

One-Sample Test						
	Test Value=4.53					
	t	df	Sig. (2-tailed)	Mean Difference	95% Confidence Interval of the Difference	
					Lower	Upper
含碳量	-1.774	8	.114	-.04000	-.0920	.0120

7.4 独立两样本 t 检验——Independent-Sample T Test 过程

单样本 t 检验是检验样本均值和总体均值是否相等。而独立两样本 t 检验是检验两个独立样本的均值是否相等。本节将通过例子介绍独立两样本 t 检验的一般步骤、操作界面及结果解释。

7.4.1 独立两样本t检验的一般步骤

独立两样本 t 检验同单样本 t 检验的步骤相类似，这里只介绍二者的区别。

- 假设不同。

在独立两样本 t 检验中， H_0 为 $\mu_1 = \mu_2$ ，即假设两样本均值相等。 H_1 为 $\mu_1 \neq \mu_2$ ，即假设两样本均值不相等。

- 统计量 t 的计算方法不同。

根据两样本的方差是否相等，将独立两样本 t 检验分为一般的 t 检验和校正的 t 检验两种情况：

(1) 当样本方差相等时，采用一般的 t 检验。此时有：

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2} \cdot \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} - t(n_1 + n_2 - 2) \quad (7.2)$$

其中 \bar{X}_i 代表样本均值， S_i 代表样本方差， n_i 代表样本个数 ($i=1,2$)。

(2) 当样本方差不相等时，采用校正的 t 检验。此时有：

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} - t(n) \quad (7.3)$$

其自由度 n 计算如下：

$$n = \left(\frac{k^2}{n_1 - 1} + \frac{(1 - k)^2}{n_2 - 1} \right)^{-1}$$

$$k = \frac{\frac{S_1^2}{n_1}}{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}$$

在进行独立两样本 t 检验之前，要通过 F 检验来看两样本的方差是否相等，从而选取恰当的统计方法。

注意 (1) 两样本必须是独立的。

(2) 样本来自的总体要服从正态分布。

7.4.2 Independent-Sample T Test过程的操作界面

执行【Analyze】/【Compare Means】/【Independent-Sample T Test】命令，弹出如图 7-6 所示对话框。

该对话框主要由以下几部分组成。

- 候选变量框

即左侧变量列表框，该变量框只显示可以进行 t 检验的变量。

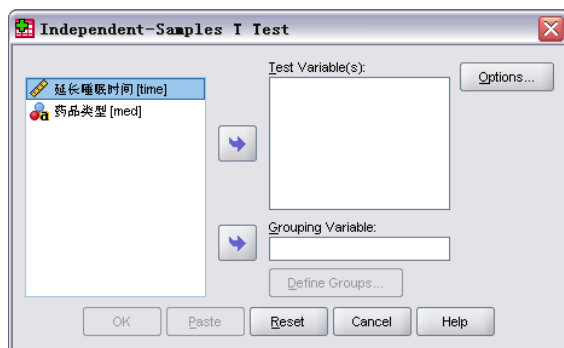


图 7-6 【Independent-Sample T Test】对话框

- 【Test Variables】

选择进行 t 检验的变量，可以同时选择多个变量。

- 【Grouping Variable】

选择分组变量。

- 【Define Groups】

定义变量的分组方法。单击【Define Groups】按钮，弹出如图 7-7 或如图 7-8 所示的对话框。

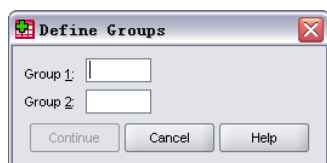


图 7-7 第一类【Define Groups】对话框

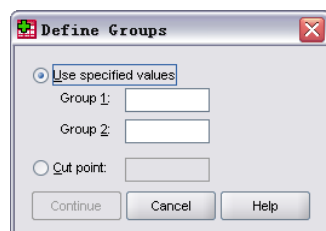


图 7-8 第二类【Define Groups】对话框

如果分组变量的测量尺度是名义型的，那么弹出如图 7-7 所示的【Define Groups】对话框。如果分组变量的测量尺度是标度型的，那么弹出如图 7-8 所示的【Define Groups】对话框。变量的分组方法主要有以下两种。

(1) Use specified values: 用特定的变量值分组。当变量的取值等于【Group1】框中自定义值时将其划为第一组；等于【Group2】框中自定义值时将其划为第二组。

(2) Cut point: 定义分割点值。当变量的取值大于或等于分割点值时将其作为第一组，小于分割点值时将其作为第二组。

- 【Options】

单击图 7-6 中的【Options】按钮，弹出如图 7-9 所示对话框。该对话框主要包括如下选项。

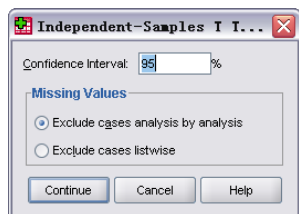


图 7-9 【Options】对话框

① Confidence Interval: 设置两样本均值之差的置信区间，默认为 95%。

② Missing Values: 设置缺失值处理方式。有两种处理方法：一种方法是仅当数据要

分析的变量值缺失时才剔除该数据 (Exclude cases analysis by analysis); 另一种方法是只要数据中有变量值缺失就剔除该数据 (Exclude cases listwise)。为了保留更多的数据样本, 默认为前者。

7.4.3 引例及结果解释

下面通过一个例子来介绍【Independent-Sample T Test】过程的操作及其结果。

例 7.3 设有甲、乙两种安眠药, 比较它们的治疗效果。以 X 表示失眠病人服用甲药后睡眠延长的时间; 用 Y 表示服用乙药后睡眠延长的时间。现在独立观察 16 个病人, 其中 8 人服甲药, 另外 8 人服乙药, 延长时间如表 7-9 所示。

表 7-9 服用甲乙药品后延长睡眠的时间

单位: (小时)

X	0.1	0.1	3.5	4.3	1.8	2.7	5.4	0.8
Y	1.7	2.2	0.0	0.6	1.5	3.3	1.5	1.2

假设 X 与 Y 都服从正态分布。试问, 这两种药的疗效有无显著差异。

这是一个独立样本的均值比较问题。要看两种药的疗效有无差异, 我们通过比较它们对病人睡眠延长时间的均值判断。执行以下操作。

STEP 01 建立数据文件 “medicine.sav”。用变量 “med” 表示样品种类, 变量 “time” 表示药品对睡眠的延长时间, 录入表 7-9 中数据。

STEP 02 进行独立两样本 t 检验, 执行以下操作:

执行【Analyze】/【Compare Means】/【Independent-Sample T Test】命令, 弹出【Independent-Sample T Test】对话框	
【Test Variables】: time	设定对睡眠的延长时间为待检验变量
【Grouping Variable】: med	设定药品类型为分组变量
单击【Define Groups】按钮	弹出【Define Groups】对话框
【Define Groups】对话框:	
【Group1】: X	定义 X 药为一组
【Group2】: Y	定义 Y 药为一组
单击【Continue】按钮	【Define Groups】对话框定义完成
单击【OK】按钮	定义完成

执行以上操作之后, 生成表 7-10 和表 7-11 两组数据。

表 7-10 所示是两种药品的延长睡眠时间统计量。包括了样本数 (N)、均值 (Mean)、标准差 (Std.Deviation) 和标准误均值 (Std.Error Mean)。

表 7-10 分组统计量

		Group Statistics			
	药品类型	N	Mean	Std. Deviation	Std. Error Mean
延长睡眠时间	X	8	2.3375	1.97769	.69922
	Y	8	1.5000	.99427	.35153

表 7-11 所示是独立两样本 t 检验结果表。这张表格给出了两种 t 检验的结果。分别为在样本方差相等（Equal Variances assumed）的情况下一般 t 检验结果和在样本方差不等（Equal Variances not assumed）的情况下的校正 t 检验结果。两种 t 检验结果到底应该选择哪一个，这取决于表 7-11 中的“Levene's Test for Equality of Variances”项，即方差齐次性检验结果。对于齐次性，这里采用的是 F 检验，由于其显著性水平 Sig. 为 0.032，小于 0.05。所以认为两样本的方差是不相等的。

在方差不相等的情况下，选取校正 t 检验方法的结果，即表 7-11 中的第二行。由于此时校正 t 检验的显著性水平 Sig. (2-tailed) 为 0.309，大于 0.05。所以不拒绝假设 H_0 ，认为两样本的均值是相等的。在本例中，X、Y 两种药品的疗效没有显著性差异。

表 7-11 独立两样本 t 检验

Independent Samples Test										
		Levene's Test for Equality of Variances		t-test for Equality of Means						
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	95% Confidence Interval of the Difference	
									Lower	Upper
延长睡眠时间	Equal variances assumed	5.637	.032	1.070	14	.303	.83750	.78261	-.84103	2.51603
	Equal variances not assumed			1.070	10.326	.309	.83750	.78261	-.89883	2.57383

注意 【Independent-Sample t Test】过程除了可以作一般的独立两样本 t 检验之外，还可以作独立两样本几何均值比较的 t 检验。该检验的操作和一般的独立两样本 t 检验操作一样。唯一的区别就是在作几何均值比较的 t 检验时，需要事先对观察值作对数变换。这可以通过【Transform】/【Compute Variable】过程来完成。

7.5 配对样本 t 检验——Paired-Sample T Test 过程

配对样本 t 检验是配对设计的样本差数的均值同总体均值 0 比较的 t 检验。本节将通过例子介绍配对样本 t 检验的一般步骤、操作界面及结果解释。

7.5.1 配对样本 t 检验一般步骤

配对样本 t 检验是针对配对数据的 t 检验。其检验方法是首先求出每对样本的差值，然后比较样本差值的均值和总体均值 0 之间的关系。如果两组数据没有差别，那么其样本差值的均值应该在 0 附近波动。否则认为两组数据是有差别的。这种方法的本质就是在对配对样本的差值同总体均值 0 作单样本 t 检验。

需要注意的是，单样本 t 检验和独立两样本 t 检验一样，样本内部数据的顺序可以任意调换。而配对样本 t 检验的样本必须是一一对应的。样本内数据的顺序不能随意交换顺序。

配对样本t检验的步骤同前面两种t检验类似，这里就只给出它们的不同之处。

(1) 假设不同：在配对样本t检验中， H_0 为 $\mu_1 = \mu_2$ ，即假设两样本均值相等。 H_1 为 $\mu_1 \neq \mu_2$ ，即假设两样本均值不相等。

(2) 检验统计量不同。在配对样本t检验中，设 x_{1i} 、 x_{2i} ($i=1\dots n$) 分别为两配对样本。其样本差值 $d_i = x_{1i} - x_{2i}$ 。此时检验统计量为：

$$t = \frac{\bar{d}}{\sqrt{S_d^2 / n}} - t(n-1) \quad (7.4)$$

其中 \bar{d} 为 d_i 的均值， S_d^2 为 d_i 的方差， n 为样本数。

从(7.4)式可以看出，配对样本t检验就是差值的单样本t检验。

7.5.2 Paired-Sample T Test过程的操作界面

执行【Analyze】/【Compare Means】/【Paired-Sample T Test】命令，弹出如图7-10所示对话框。

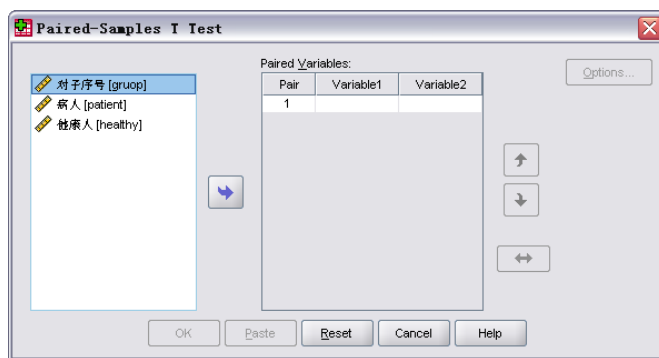


图7-10 【Paired-Sample T Test】对话框

该对话框主要由以下几部分组成。

- 候选变量框

即左侧变量列表框，该变量框内显示可以进行配对样本t检验的变量。

- 【Paired Variables】

选择进行t检验的配对样本。将两个变量分别移入 Variable1 栏和 Variable2 栏下边，系统自动将这两个变量配对。可以同时选入多对配对样本。

- 【Options】

单击【Options】按钮，弹出如图7-11所示对话框。

该对话框主要包括以下几项。

- Confidence Interval: 设置样本差的均值和总体均值 0 之间的置信区间，默认为 95%。

- Missing Values: 设置缺失值处理方式。有两种处理方法：一种是只有要分析的变量值缺失时才剔除该数据 (Exclude cases analysis by analysis)；另一种是只要数据中有变量

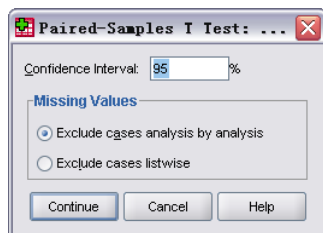


图7-11 【Options】对话框

值缺失就剔除该数据（Exclude cases listwise）。为了保留更多的数据样本，默认为前者。

7.5.3 引例及结果解释

下面通过一个例子来介绍【Paired-Sample T Test】过程的操作及其结果。

例 7.4 慢性支气管炎病人血液中胆碱酯酶活性常常偏高。某高校将同性别同年龄的病人与健康人配成 8 对，测量值如表 7-12 所示。根据测量值能否得出结论，即病人血液中胆碱酯酶活性的确比健康人偏高。

表 7-12 慢性气管炎病人与健康人血液胆碱酯酶活性测定（ $\mu\text{M}/\text{ml}$ ）

对子序号	病人组	健康人组
1	3.28	2.36
2	2.60	2.40
3	3.32	2.40
4	2.72	2.52
5	2.38	3.04
6	3.64	2.64
7	2.98	2.56
8	4.40	2.40

通过配对样本 t 检验的方法来看病人与正常人的血液中胆碱酯酶活性是否有明显差别。操作如下：

STEP 01 建立数据文件“blood.sav”。用变量“group”表示配对样本的对子序号，变量“patient”表示病人血液中胆碱酯酶活性的测量值，变量“healthy”表示健康人血液中胆碱酯酶活性的测量值。录入表 7-12 中的数据。

STEP 02 配对样本 t 检验，执行以下操作：

执行【Analyze】/【Compare Means】/【Paired-Sample T Test】命令，弹出【Paired-Sample T Test】对话框

【Paired Variables】：patient - healthy 设定待检验配对样本

单击【OK】按钮 定义完成

STEP 03 执行以上操作之后，生成表 7-13～表 7-15。

表 7-13 是病人和健康人血液中胆碱酯酶活性的测量值的分组统计量。包括均值（Mean）、样本数（N）、标准差（Std.Deviation）、标准误均值（Std.Error Mean）。从表 7-13 可以看出，病人血液中胆碱酯酶活性的测量值的均值偏大且扰动也偏大。

表 7-13 分组统计量

Paired Samples Statistics					
		Mean	N	Std. Deviation	Std. Error Mean
Pair	病人	3.1650	8	.64981	.22974
1	健康人	2.5400	8	.22424	.07928

表 7-14 是配对样本的相关性分析结果。其相关系数为-0.467，对应的显著性水平 Sig. 为 0.224，大于 0.05，即认为两样本的相关性不显著。

表 7-14 配对样本的相关性分析

Paired Samples Correlations			
	N	Correlation	Sig.
Pair 1 病人和健康人	8	-.467	.244

表 7-15 所示是配对样本 t 检验结果。其显著性水平 Sig.取值大于 0.05，即认为病人和健康人的血液中胆碱酯酶活性的测量值没有显著的差别。

表 7-15 配对样本 t 检验

Paired Samples Test								
	Paired Differences					t	df	Sig. (2-tailed)
	Mean	Std. Deviation	Std. Error	95% Confidence Interval of the Difference				
				Lower	Upper			
Pair 1 病人-健康人	.62500	.78009	.27580	−.02717	1.27717	2.266	7	.058

7.6 本章小结

本章介绍了 SPSS 均值比较和 t 检验的【Compare Means】子菜单，详细介绍了以下几个过程：

- Means 过程，分组计算样本的描述性统计量；
- One-Sample T Test 过程，单样本 t 检验；
- Independent-Sample T Test 过程，独立两样本 t 检验；
- Paired-Sample T Test 过程，配对样本 t 检验。

t 检验主要用来处理两样本间均值比较的问题，这些过程在实际问题中经常用到，读者务必掌握。同时，t 检验只是一类特殊的假设检验，通过本章的学习，读者可以了解假设检验的一般步骤和基本原理。

第 8 章 方差分析

上一章介绍了 SPSS 中 t 检验的实现过程。 t 检验解决了两样本间均值比较的问题。对于多个总体均值的检验，需要引入方差分析。本章通过例子详细介绍各类常见方差分析方法及其在 SPSS 中的实现。本章内容包括：

- 方差分析简介
- 单因素方差分析——One-Way ANOVA 过程
- 多因素方差分析——Univariate 过程（1）
- 协方差分析——Univariate 过程（2）

8.1 方差分析简介

本节主要概括性地介绍方差分析的概念、应用背景和分类等。各类具体的方差分析方法将在本章的后续几节详细介绍。

8.1.1 方差分析的提出

在科学试验和生产过程中，影响一事物的因素是多方面的。比如农作物的产量受到品种、肥料、水分和气候等因素的影响。这些因素有的对产量影响大一些，有的对产量影响小一些。那么对于产量，究竟哪些因素的影响是显著的，哪些因素的影响是不显著的呢？同时，除了上述因素之外，产量的差异还要受随机误差的影响。随机误差对产量的影响到底有多大？这些问题都可以通过方差分析来解决。方差分析就是采用数理统计的方法对所有结果进行分析，是鉴别各种因素对研究对象的某些特征值影响大小的一种有效的方法。

方差分析最早是由 R.A.Fisher 于 1920 年前后对农业试验作统计分析时提出的。由于它可以由较少的试验有效地获得大量的信息，所以其应用范围从最初的生物农业扩大到现在的各个领域。

8.1.2 方差分析的基本概念

下面通过一个简单的例子来介绍方差分析中的基本概念。

例 8.1 为了寻求适应某地区的高产油菜品种。现选了 5 种不同品种进行试验，每一品种在 4 块条件完全相同的试验田上试种，其他施肥等田间管理措施完全一样。表 8-1 所

示为每一品种下每一块田的亩产量和每一品种下 4 块田的平均亩产量。（数据来源：《常用统计方法》，华东师范大学出版社）

表 8-1 油菜产量数据

田 块 \ 品 种	A_1	A_2	A_3	A_4	A_5
1	256	244	250	288	206
2	222	300	277	280	212
3	280	290	230	315	220
4	298	275	322	259	212
平均亩产	264	277.25	269.75	285.50	212.50

根据这些数据分析不同油菜品种对平均亩产影响是否显著。


在方差分析中，主要有以下几个常用概念。

(1) 试验指标：把研究对象的特征值，即试验结果称为试验指标，简称为指标。通常用 X 表示。在本例中是指油菜的产量。

(2) 因素：指可能对试验指标产生影响的变量。通常用大写字母 A 、 B 、 C 、 D 等表示。例 8.1 中只有一个因素即油菜的品种。

(3) 水平：因素的不同状态。本例中因素有 5 个水平，分别为 A_1 、 A_2 、 A_3 、 A_4 、 A_5 。

例 8.1 的问题就是辨别油菜产量的差异主要是由抽样误差造成的还是由油菜品种不同造成的。这一问题可以归结为判断 5 个总体是否具有相同的分布。在安排试验的时候，除了油菜的品种，其他的试验条件总是尽可能做到一致，这使我们可以认为总体的方差是相同的。因此在这里，推断几个总体是否具有相同分布的问题就简化为检验这个具有相同方差的总体其均值是否相等的问题。

 **注意** 通过上面的的分析可以得出方差分析的适用条件。

- (1) 样本来自的总体要服从正态分布。
- (2) 样本方差必须是齐次的。
- (3) 各样本之间相互独立。

8.1.3 方差分析的类型

根据试验指标和影响指标的因素个数，将常见的方差分析分为如表 8-2 所示的几类。

表 8-2 方差分析的类型

名 称	指标个数	因素个数	SPSS 中的实现
单因素方差分析	1 个	1 个	执行【Analyze】/【Compare Means】/【One-Way ANOVA】命令
多因素方差分析	1 个	2 个或 2 个以上	执行【Analyze】/【General Linear Model】/【Univariate】命令
多元方差分析	2 个及 2 个以上	1 个或多个	执行【Analyze】/【General Linear Model】/【Multivariate】命令

除了表 8-2 中所提到的一般方差分析之外，还有一类特殊的方差分析即协方差分析。协方差分析是一种将一般的方差分析和回归分析结合起来的统计方法。与多因素方差分析

一样，它也是通过【General Linear Model】菜单下的【Univariate】过程来实现的。

在本章的后几节，就将通过具体的例子来学习单因素方差分析、多因素方差分析和协方差分析。对于多元方差分析，由于 SPSS【Multivariate】过程和【Univariate】过程的操作界面几乎完全一致，所以此处不再复述。

8.2 单因素方差分析——One-Way ANOVA过程

虽然是最简单的方差分析，但单因素方差分析在实际问题中还是有着广泛的应用。本节将通过例子来介绍单因素方差分析的一般步骤、操作界面及结果解释。

8.2.1 单因素方差分析简介

单因素方差分析是指只单独考虑一个因素 A 对指标 X 的影响。此时其他因素都不变或者控制在一定的范围之内。考虑因素 A 有 k 个水平，在每个水平下做 n_i 次试验，则有如表 8-3 所示的试验数据。

表 8-3 单因素方差分析原始数据表

水 平 重 复	A_1	A_2	...	A_k
1	x_{11}	x_{21}	...	x_{k1}
2	x_{12}	x_{22}	...	x_{k2}
...
n_i	x_{1n_i}	x_{2n_i}	...	x_{kn_i}

在介绍单因素方差分析之前，首先引入如下记号。

① 试验总次数： $n = n_1 + n_2 + \dots + n_k$

② 水平 i 下的样本均值： $\bar{x}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} x_{ij}$

③ 总体样本均值： $\bar{x} = \frac{1}{n} \sum_{i=1}^k \sum_{j=1}^{n_i} x_{ij} = \frac{1}{k} \sum_{i=1}^k \bar{x}_i$

④ 总离差平方和： $S_T = \sum_{i=1}^k \sum_{j=1}^{n_i} (x_{ij} - \bar{x})^2$

⑤ 因素 A 的组间平方和： $S_A = \sum_{i=1}^k n_i (\bar{x}_i - \bar{x})^2$ ，它反映了因素 A 的水平差异对指标所产生的影响。

⑥ 误差平方和： $S_e = \sum_{i=1}^k \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_i)^2$ ，它反映了随机误差及其他因素对指标所产生的综合影响，也称为组内平方和。

通过简单的计算，有

$$S_T = S_A + S_e \quad (8.1)$$

(8.1) 式表明, 总离差平方和可以分解为因素 A 的组间平方和与误差平方和两部分。这就是方差分析的本质所在。现在的问题是研究因素 A 对指标 X 的影响是否显著。由(8.1)式可知, 当 S_T 一定时, 若因素 A 对指标 X 有显著影响, 则 S_A 越大, S_e 越小。否则, 则 S_A 越小, S_e 越大。因此, 构造如下统计量:

$$F = \frac{\frac{S_A}{k-1}}{\frac{S_e}{n-k}} \quad (8.2)$$

由数理统计知识可知, F 服从自由度为 $(k-1, n-k)$ 的 F 分布。从前面的分析可以发现, 因素 A 对指标 X 的影响越显著, 则 F 值越大。以下是单因素方差分析的一般步骤。

STEP 01 提出假设: 认为因素 A 对指标 X 的影响不显著, 即可以认为 x_{ij} 来自同一总体, 则有 $H_0: \mu_1 = \mu_2 = \dots = \mu_k$ 。

STEP 02 选取统计量 F 作为检验统计量, 并确定其分布。

STEP 03 选择显著性水平 α , 并根据统计量 F 的分布, 确定临界值 F_0 及 H_0 的拒绝域。

STEP 04 例如表 8-4 所示的方差分析表, 计算检验统计量 F。

表 8-4 单因素方差分析表

方差来源	平方和	自由度	均方和	F 值	F 临界值
因素 A	S_A	$k-1$	$V_A = S_A / (k-1)$	$F = \frac{\frac{S_A}{k-1}}{\frac{S_e}{n-k}}$	$F_{\alpha}(k-1, n-k)$
误差	S_e	$n-k$	$V_e = S_e / (n-k)$		
总和	S_T	$n-1$			

STEP 05 下结论: 一般地, 若 $F_{0.05}(k-1, n-k) < F$, $F_{0.01}(k-1, n-k)$, 则称因素 A 对指标 X 影响显著。若 $F > F_{0.01}(k-1, n-k)$, 则称因素 A 对指标 X 的影响高度显著。

8.2.2 One-Way ANOVA过程的操作界面

首先介绍单因素方差分析的操作界面。执行【Analyze】/【Compare Means】/【One-Way ANOVA】命令, 弹出如图 8-1 所示对话框。

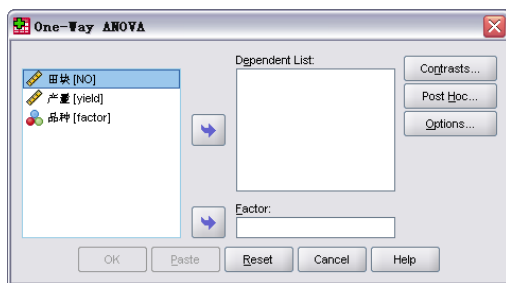


图 8-1 【One-Way ANOVA】对话框

该对话框主要由以下几部分组成。

1. 候选变量框

即左侧变量列表框。

2. 【Dependent List】

选择单因素方差分析的指标变量。可以同时选择多个变量，此时 SPSS 就将分别对各指标作单因素方差分析。

3. 【Factor】

选择因素变量。由于进行的是单因素方差分析，所以此时只能选择一个变量。

4. 【Contrasts】

单击【Contrasts】按钮，弹出如图 8-2 所示对话框。

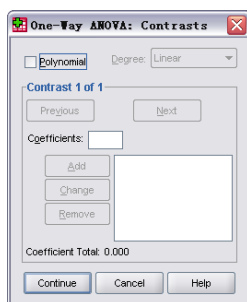


图 8-2 【Contrasts】对话框

该对话框主要用于对组间平方和进行分解并确定均值的多项式比较。主要包括以下几项。

① Polynomial 复选框：选择是否对方差分析的组间平方和进行分解并进行趋势检验。

② Degree 下拉列表：选中 Polynomial 复选框后，该下拉列表被激活。用于选择进行趋势检验的曲线类型。

③ Coefficients 框：精确定义均值比较的多项式系数。比如有 A1、A2、A3、A4 四个因素水平，想要比较 A1 和 A4 水平下的样本均值。那么就在 Coefficients 框输入“1”，然后单击【Add】按钮将其添加到下方框中；再按照相同的方法分别添加“0”、“0”、“-1”到框中。这样，在 SPSS 的结果输出窗口就会出现想要的检验结果。需要注意的是，所有添加的系数加起来必须等于 0。同时可以通过【Next】按钮定义多组这样的多项式系数。

5. 【Post Hoc】

单击图 8-1 中的【Post Hoc】按钮，弹出如图 8-3 所示对话框。

该对话框主要用于定义多重比较的检验方法。比如，方差分析的结果认为因素 A 各水平之间的差异会对指标 X 造成显著影响。但是这并不意味着任意两个水平之间的差异都会给指标 X 造成显著影响。要解决这个问题，就有必要将各个水平的均值进行两两比较，这种两两比较的方法就称为多重比较。该对话框主要包括以下几项。

① Equal Variances Assumed 复选框组：该组主要用于定义在样本方差齐次的情况下多重比较的检验方法。SPSS 共提供了 14 种检验方法。其中最常用的是 LSD 法和 S-N-K 法。

② Equal Variances Not Assumed 复选框组：该组主要用于定义在样本方差不齐次的情况下多重比较的检验方法。

③ Significance Level 框：定义两两比较的显著性水平。

6. 【Options】

单击图 8-1 上的【Options】按钮，弹出如图 8-4 所示的对话框。

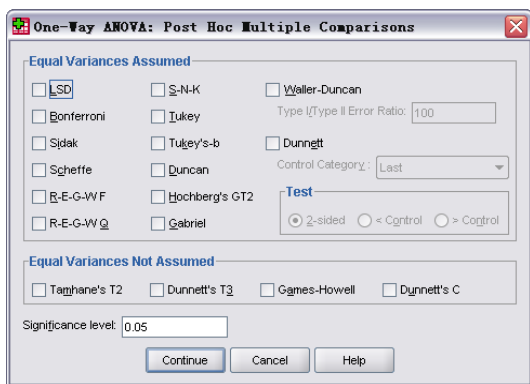


图 8-3 【Post Hoc】对话框

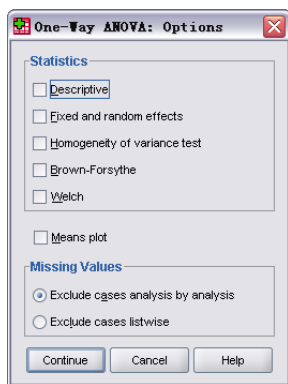


图 8-4 【Options】对话框

该对话框主要包括以下选项。

• Statistics 复选框组

定义可选统计指标。主要包括以下几类统计指标。

① Descriptives：输出各组的描述性统计量。

② Fixed and random effects：输出不变效应模型和随机效应模型的标准差、标准误差，以及 95% 的置信区间。

③ Homogeneity of variance test：计算 Levene 统计量，检验各组的方差齐次性。

④ Brown-Forsythe：计算 Brown-Forsythe 统计量，检验各组的均值是否相等。在方差不齐次的情况下，此方法比方差分析可靠。

⑤ Welch：计算 Welch 统计量，检验各组的均值是否相等。在方差不齐次的情况下，这种方法也比方差分析可靠。

• Means plot 复选框

定义是否绘制各组均值的图形，以便从图形上直观比较各组间均值的差异。

• Missing Values 单选框组

定义缺失值的处理方式。

8.2.3 引例及结果解释

通过例 8.1 来介绍【One-Way ANOVA】过程的操作及其结果。对例 8.1 执行以下操作。

STEP 01 建立如图 8-5 所示的数据文件“yield.sav”。其中“NO”表示同一水平下的

第 i 次试验。“yield”表示该次试验下油菜产量，“factor”表示油菜品种。

	NO	yield	factor
1	1	256.00	1
2	2	222.00	1
3	3	260.00	1
4	4	298.00	1
5	1	244.00	2

图 8-5 数据文件“yield.sav”的数据结构

STEP 02 进行单因素方差分析，执行以下操作：

执行【Analyze】/【Compare Means】/【One-Way ANOVA】命令，弹出【One-Way ANOVA】对话框	
【Dependent List】：yield	定义油菜产量为指标变量
【Factor】：factor	定义油菜品种为因素变量
单击【Post Hoc】按钮	弹出【Post Hoc】对话框
【Post Hoc】对话框：	
选中“LSD”复选框	定义用 LSD 法进行多重比较检验
单击【Continue】按钮	【Post Hoc】对话框定义完成
单击【Options】按钮	弹出【Options】对话框
【Options】对话框：	
选中“Homogeneity of variance test”复选框	定义检验各组的方差齐次性
选中“Means plot”复选框	定义绘制均值的图形
单击【Continue】按钮	【Options】对话框定义完成
单击【OK】按钮	定义完成

执行以上操作之后，生成表 8-5～表 8-7。

表 8-5 是各组数据的方差齐次性检验结果。由于表中计算 Levene 统计量取值为 1.896，Sig. 值为 0.164，大于 0.05，所以认为各组的方差齐次。

表 8-5 方差齐次性检验结果

Test of Homogeneity of Variances			
Levene Statistic	df 1	df 2	Sig.
1.896	4	15	.164

表 8-6 是一个典型的方差分析表。从左到右分别为平方和（Sum of Squares）、自由度（df）、均方和（Mean Square）、F 值和显著性指标（Sig.）。从表 8-6 中数据可知，Sig. 取值小于 0.05，即假设不成立，认为各组的均值是有差异的。也就是说至少有一类油菜品种的产量和其他品种有显著差异。

表 8-6 方差分析表

ANOVA

产量

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	13195.700	4	3298.925	4.306	.016
Within Groups	11491.500	15	766.100		
Total	24687.200	19			

表 8-7 是各类油菜品种之间显著性差异两两比较的结果。从表中数据可以看出, 由于品种 5 同其他任意 4 个品种比较其 Sig.取值都是小于 0.05 的, 所以认为其同其他品种在产量上有显著差异。而另外 4 个品种的油菜可以认为其品种的不同对产量的影响不显著。

表 8-7 多重检验表

Multiple Comparisons

Dependent Variable: 产量

LSD

(I)品种	(J)品种	Mean Difference (I-J)	Std Error	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
1	2	-13.25000	19.57166	.509	-54.9660	28.4660
	3	-5.75000	19.57166	.773	-47.4660	35.9660
	4	-21.50000	19.57166	.289	-63.2160	20.2160
	5	51.50000*	19.57166	.019	9.7840	93.2160
2	1	13.25000	19.57166	.509	-28.4660	54.9660
	3	7.50000	19.57166	.707	-34.2160	49.2160
	4	-8.25000	19.57166	.679	-49.9660	33.4660
	5	64.75000*	19.57166	.005	23.0340	106.4660
3	1	5.75000	19.57166	.773	-35.9660	47.4660
	2	-7.50000	19.57166	.707	-49.2160	34.2160
	4	-15.75000	19.57166	.434	-57.4660	25.9660
	5	57.25000*	19.57166	.010	15.5340	98.9660
4	1	21.50000	19.57166	.289	-20.2160	63.2160
	2	8.25000	19.57166	.679	-33.4660	49.9660
	3	15.75000	19.57166	.434	-25.9660	57.4660
	5	73.00000*	19.57166	.002	31.2840	114.7160
5	1	-51.50000*	19.57166	.019	-93.2160	-9.7840
	2	-64.75000*	19.57166	.005	-106.4660	-23.0340
	3	-57.25000*	19.57166	.010	-98.9660	-15.5340
	4	-73.00000*	19.57166	.002	-114.7160	-31.2840

*. The mean difference is significant at the .05 level.

图 8-6 所示是各组间均值比较的线图。从图形上也正好可以印证表 8-7 的结论。通过观察图形可以发现, 品种 5 的平均产量远低于其他品种。

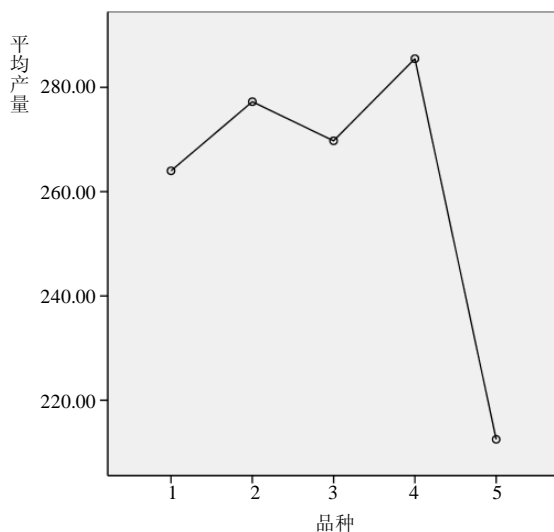


图 8-6 各组均值比较图

8.3 多因素方差分析——Univariate过程（1）

上一节介绍了单因素方差分析，然而客观世界是十分复杂的，影响一个指标的因素也是多方面的。因此，在实际问题中又引入了多因素方差分析的方法。本节将通过例子介绍多因素方差分析的一般步骤、操作界面及结果解释。

8.3.1 多因素方差分析简介

多因素方差分析考虑的是有多个因素同时对指标产生影响的问题。这些因素之间往往又相互联系、相互影响。随着因素增多，问题也变得更加复杂。这里为了简化问题，只考虑两因素方差分析。根据两因素联合作用是否会对指标产生显著影响，将其分为无交互作用的两因素方差分析和有交互作用的两因素方差分析。

1. 无交互作用的两因素方差分析

所谓无交互作用是指两因素 A、B 的联合作用不会对指标 X 形成显著影响。对于指标 X，设有两因素 A、B，因素 A 有 r 个水平 $A_1、A_2、\cdots、A_r$ ，因素 B 有 s 个水平 $B_1、B_2、\cdots、B_s$ 。现在考虑在两因素无交互作用的情况下因素 A、B 是否会对指标 X 造成显著影响。对于因素 A、B 各个水平的每一对组合 (A_i, B_j) 都只进行一次相互独立的试验，试验结果如表 8-8 所示。

与单因素方差分析类似，两因素方差分析的基本思想还是通过分析造成指标 X 总离差平方和的原因来确定各因素对指标的影响是否显著。首先引入如下记号。

表 8-8 无交互作用的两因素方差分析原始数据表

因素 A \ 因素 B	B_1	...	B_j	...	B_s
A_1	x_{11}	...	x_{1j}	...	x_{1s}
...
A_i	x_{i1}	...	x_{ij}	...	x_{is}
...
A_r	x_{r1}	...	x_{rj}	...	x_{rs}

① 试验总次数: $n = r \times s$

② 水平 A_i 下的样本均值: $\bar{x}_{i.} = \frac{1}{s} \sum_{j=1}^s x_{ij}$

③ 水平 B_j 下的样本均值: $\bar{x}_{.j} = \frac{1}{r} \sum_{i=1}^r x_{ij}$

④ 总体样本均值: $\bar{x} = \frac{1}{n} \sum_{i=1}^r \sum_{j=1}^s x_{ij}$

⑤ 总离差平方和: $S_T = \sum_{i=1}^r \sum_{j=1}^s (x_{ij} - \bar{x})^2$

⑥ 因素 A 的组间平方和: $S_A = \sum_{i=1}^r s(\bar{x}_{i.} - \bar{x})^2$, 它反映了因素 A 的水平差异对指标所产生的影响。

⑦ 因素 B 的组间平方和: $S_B = \sum_{j=1}^s r(\bar{x}_{.j} - \bar{x})^2$, 它反映了因素 B 的水平差异对指标所产生的影响。

⑧ 误差平方和: $S_e = \sum_{i=1}^r \sum_{j=1}^s (x_{ij} - \bar{x}_{i.} - \bar{x}_{.j} + \bar{x})^2$, 它反映了随机误差及其他因素对指标所产生的综合影响。

通过简单的计算, 有

$$S_T = S_A + S_B + S_e \quad (8.3)$$

与单因素方差分析类似, 还是可以通过构造 F 统计量的方法来检验因素 A、B 对指标的影响是否显著。具体过程这里不再复述。表 8-9 是无交互作用的两因素方差分析表。

表 8-9 无交互作用的两因素方差分析表

方差来源	平方和	自由度	均方和	F 值	F 临界值
因素 A	S_A	$f_A = r - 1$	$V_A = S_A / f_A$	$F_A = \frac{S_A / f_A}{S_e / f_e}$	$F_{\alpha}(r-1, (r-1)(s-1))$
因素 B	S_B	$f_B = s - 1$	$V_B = S_B / f_B$	$F_B = \frac{S_B / f_B}{S_e / f_e}$	$F_{\alpha}(s-1, (r-1)(s-1))$
误差	S_e	$f_e = (r-1)(s-1)$	$V_e = S_e / f_e$		
总和	S_T	$rs - 1$			

一般地，若 $F_{0.05}(r-1, (r-1)(s-1)) < F_A$ ， $F_{0.01}(r-1, (r-1)(s-1))$ ，则称因素 A 对指标 X 影响显著。若 $F_A > F_{0.01}(r-1, (r-1)(s-1))$ ，则称因素 A 对指标 X 的影响高度显著。对于因素 B，也可以由 F_B 做出类似的结论。

2. 有交互作用的两因素方差分析

前面讨论的问题是假设指标 X 的两个因素的联合作用不会对指标造成显著影响。现在讨论当因素 A、B 联合作用会对指标造成显著影响时，如何通过试验结果来判断各因素及其交互作用对指标的影响是否显著。

对于因素 A、B 各个水平的每一对组合 (A_i, B_j) 都进行 m 次试验，试验结果如表 8-10 所示。

表 8-10 有交互作用的两因素方差分析原始数据表

因素 A \ 因素 B	B_1	...	B_j	...	B_s
A_1	$x_{111} \dots x_{11m}$...	$x_{1j1} \dots x_{1jm}$...	$x_{1s1} \dots x_{1sm}$
...
A_i	$x_{i11} \dots x_{i1m}$...	$x_{ij1} \dots x_{ijm}$...	$x_{is1} \dots x_{ism}$
...
A_r	$x_{r11} \dots x_{r1m}$...	$x_{rj1} \dots x_{rjm}$...	$x_{rs1} \dots x_{rsm}$

首先引入如下记号。

① 试验总次数： $n = r \times s \times m$

② 水平 A_i 下的样本均值： $\bar{x}_{i..} = \frac{1}{m \times s} \sum_{j=1}^s \sum_{k=1}^m x_{ijk}$

③ 水平 B_j 下的样本均值： $\bar{x}_{.j.} = \frac{1}{m \times r} \sum_{i=1}^r \sum_{k=1}^m x_{ijk}$

④ 水平 $A_i B_j$ 下的样本均值： $\bar{x}_{ij.} = \frac{1}{m} \sum_{k=1}^m x_{ijk}$

⑤ 总体样本均值： $\bar{x} = \frac{1}{n} \sum_{i=1}^r \sum_{j=1}^s \sum_{k=1}^m x_{ijk}$

⑥ 总离差平方和： $S_T = \sum_{i=1}^r \sum_{j=1}^s \sum_{k=1}^m (x_{ijk} - \bar{x})^2$

⑦ 因素 A 的组间平方和： $S_A = \sum_{i=1}^r sm(\bar{x}_{i..} - \bar{x})^2$ ，它反映了因素 A 的水平差异对指标所产生的影响。

⑧ 因素 B 的组间平方和： $S_B = \sum_{j=1}^s rm(\bar{x}_{.j.} - \bar{x})^2$ ，它反映了因素 B 的水平差异对指标所产生的影响。

⑨ 因素 A 与 B 交互作用的离差平方和： $S_{A \times B} = \sum_{i=1}^r \sum_{j=1}^s m(\bar{x}_{ij.} - \bar{x}_{i..} - \bar{x}_{.j.} + \bar{x})^2$ ，它反映了 A、B 交互作用的显著性水平。

⑩ 误差平方和： $S_e = \sum_{i=1}^r \sum_{j=1}^s \sum_{k=1}^m (x_{ijk} - \bar{x}_{ij.})^2$ ，它反映了随机误差及其他因素对指标所产生的综合影响。

通过简单的计算，有

$$S_T = S_A + S_B + S_{A \times B} + S_e \quad (8.4)$$

仍然通过构造 F 统计量的方法来检验因素 A、B 及其交互作用对指标的影响是否显著。

表 8-11 所示是有交互作用的两因素方差分析表。

表 8-11 有交互作用的两因素方差分析表

方差来源	平方和	自由度	均方和	F 值	F 临界值
因素 A	S_A	$f_A = r - 1$	$V_A = S_A / f_A$	$F_A = \frac{S_A / f_A}{S_e / f_e}$	$F_{\alpha}(f_A, f_e)$
因素 B	S_B	$f_B = s - 1$	$V_B = S_B / f_B$	$F_B = \frac{S_B / f_B}{S_e / f_e}$	$F_{\alpha}(f_B, f_e)$
A×B	$S_{A \times B}$	$f_{A \times B} = (r - 1)(s - 1)$	$V_{A \times B} = S_{A \times B} / f_{A \times B}$	$F_{A \times B} = \frac{S_{A \times B} / f_{A \times B}}{S_e / f_e}$	$F_{\alpha}(f_{A \times B}, f_e)$
误差	S_e	$f_e = rs(m - 1)$	$V_e = S_e / f_e$		
总和	S_T	$rsm - 1$			

一般地，若 $F_{0.05}(r - 1, rs(m - 1)) < F_A$ ， $F_{0.01}(r - 1, rs(m - 1))$ ，则称因素 A 对指标 X 影响显著。若 $F_A > F_{0.01}(r - 1, rs(m - 1))$ ，则称因素 A 对指标 X 的影响高度显著。对于因素 B，也可以由 F_B 做出类似的结论。若 $F_{A \times B} > F_{0.05}((r - 1)(s - 1), rs(m - 1))$ ，则说明因素 A、B 对指标有显著的交互影响。

对于因素个数大于 2 的方差分析问题，其基本思想与两因素方差分析问题是类似的。但是随着因素个数的增多，如何进行试验设计也是一个重要的问题，有兴趣的读者可以自行查阅相关书籍。

8.3.2 Univariate过程的操作界面

首先介绍【Univariate】过程的操作界面。执行【Analyze】/【General Linear Model】/【Univariate】命令，弹出如图 8-7 所示对话框。

该对话框主要由以下几部分构成。

1. 【Dependent Variable】

选择指标变量。

2. 【Fixed Factor】

选择固定因素变量。

3. 【Random Factor】

选择随机因素变量。

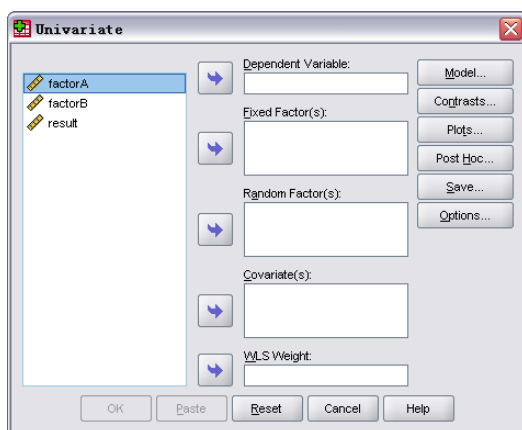


图 8-7 【Univariate】对话框

4. 【Covariate】

选择协变量，此项功能将在下一节协方差分析中用到。

5. 【WLS Weight】

选择加权最小二乘法的权重系数。

• 【Model】

单击该按钮，弹出如图 8-8 所示【Model】对话框。该对话框主要用来定义方差分析的模型。

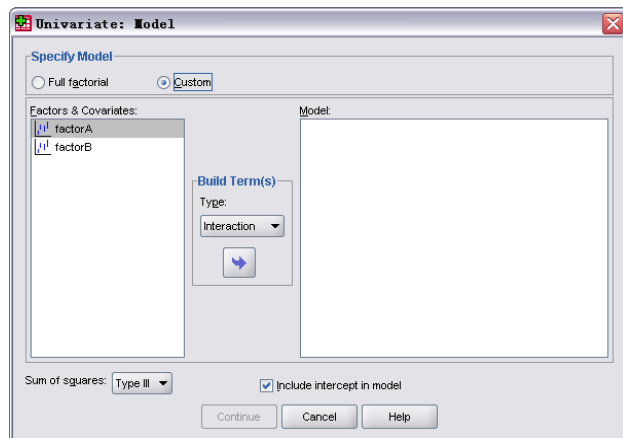


图 8-8 【Model】对话框

【Model】对话框主要包含以下几部分。

① Full factorial 框：系统默认选项。选择该项，则建立全模型，分析所有因素的主效应及其交互效应。

② Custom 框：用户自定义方差分析的模型。选择该项，则激活下方的两个选项框和一个下拉列表，这 3 项功能如表 8-12 所示。

表 8-12 Custom 选项组各部分功能

名 称	选 项	功 能
Factors&Covariates 框	--	列出了在【Univariate】过程中选择的所有的固定因素变量（F）、随机因素变量（R）和协变量（C）
Model 框	--	选择方差分析的主效应项。若同时将 Factors&Covariates 框中两个变量选入，则将其交互效应强行纳入模型
Build Term 下拉列表	Interaction	定义进行选择变量的交互效应的方差分析
	Main effects	定义进行选择变量的主效应的方差分析。若因素间无交互作用则选择此项
	All 2-way—All 5-way	定义进行所有变量的 i 阶交互效应的方差分析

③ Sum of squares 下拉列表：定义平方和的分解方式，此时一般默认选择 Type III。

④ Include intercept in model 复选框：用于选择模型中是否包含截距平方和。

6. 【Contrasts】

单击图 8-7 中【Univariate】对话框中的【Contrasts】按钮，弹出如图 8-9 所示的【Contrasts】对话框。

该对话框主要用于比较各因素水平之间的差异。包括以下几项。

• Factors 框

列出所有因素变量。变量名后括号内的是当前所选的比较方法。

• Contrast 下拉列表

设置比较因素水平间差异的方法，各方法的具体解释如下。

① None：默认选项，不比较。

② Deviation：偏差比较法。比较因素在各水平下的均值和总体均值的差异。

③ Simple：简单比较法。以最后一个或第一个因素水平下的均值为标准，比较其他水平下样本均值与其差异。

④ Difference：向前差异比较法。除第一个水平之外，因素在每个水平下的均值都与其前面所有水平的均值比较。

⑤ Helmert：向后差异比较法。与 Difference 恰好相反，除最后一个水平外，因素在每个水平下的均值都和其后面所有水平的均值比较。

⑥ Repeated：邻近比较法。除第一个水平之外，因素在每个水平的均值都和其前一个水平均值相比较。

⑦ Polynomial：多项式比较法。在有 n 个水平的情况下，比较其从 1 到 $n-1$ 次方的效应。

• Change 按钮

单击该按钮，改变选中因素的比较方法。

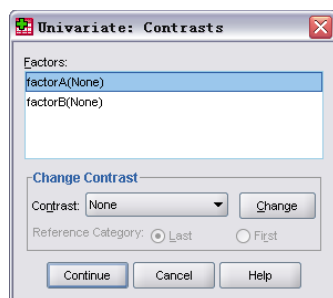


图 8-9 【Contrasts】对话框

7. 【Plots】

单击如图 8-7 所示的【Plots】按钮，弹出如图 8-10 所示对话框。该对话框主要用于定义输出的图形。

- ① Factors 框：列举可用于作图的变量。
- ② Horizontal Axis 框：选择作为横坐标的因素变量。
- ③ Separate Lines 框：选择曲线分组变量，按照该变量的不同取值在同一张图上绘制多条曲线。
- ④ Separate Plots 框：选择图形分组变量，按照该变量的不同取值绘制多张图形。
- ⑤ Plots 按钮组：用于添加（Add）、改变（Change）和移出（Remove）已定义图形。

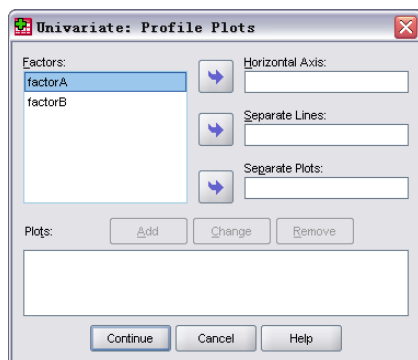


图 8-10 【Plots】对话框

8. 【Post Hoc】

单击【Post Hoc】按钮，弹出如图 8-11 所示的【Post Hoc】对话框。该对话框与单因素方差分析的【Post Hoc】对话框（如图 8-3 所示）完全类似，主要用于定义各因素的多重比较的检验方法，这里不再重复。

9. 【Save】

单击图 8-7 中的【Save】按钮，弹出如图 8-12 所示对话框。

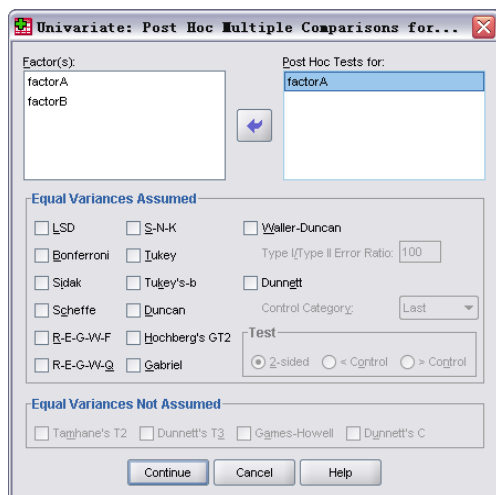


图 8-11 【Post Hoc】对话框

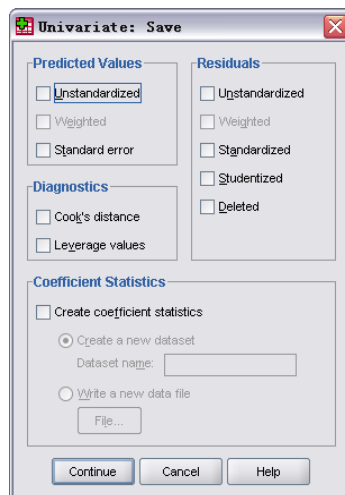


图 8-12 【Save】对话框

该对话框主要用于定义如表 8-13 所示的可保存结果，其具体的解释将在第 10 章回归分析中介绍。最后在原始的“*.sav”数据文件中新生成一列或几列新变量来保存这些选中的中间结果。

表 8-13 Save 对话框可以保存的结果

SPSS 中表示	名 称	选 项
Predicted Values	预测值	Unstandardized (未标准化预测值)
		Weighted (加权预测值, 该项仅当有 WLS 变量时才可用)
		Standard error (未标准化预测值的标准误差)
Diagnostics	诊断方法	Cook's distance (库克距离)
		Leverage values (非中心化杠杆值)
Residuals	残差	Unstandardized (未标准化残差)
		Weighted (加权残差)
		standardized (标准化残差)
		Studentized (学生残差)
		Deleted (剔除残差)

【Save】对话框上还有一个 Coefficient Statistics 组。该组主要用于保存参数拟合的协方差矩阵。SPSS 17.0 提供了两种保存方法，可以将结果保存在一个新生成的“*.sav”数据文件中(create a new data set)，也可以将结果直接保存到其他文件里(write a new data file)。

10. 【Options】

单击图 8-7 中的【Options】按钮，弹出如图 8-13 所示对话框。该对话框主要由以下几部分组成。

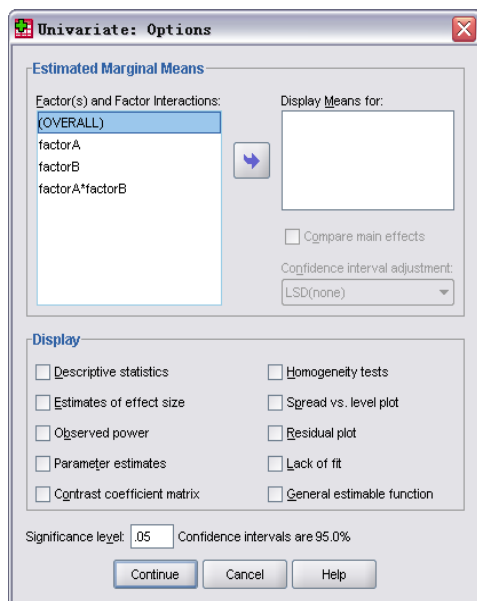


图 8-13 【Options】对话框

- Factor and Factor interactions 框

列出可选的因素变量及其交互作用。其中 OVERALL 代表对所有的因素及其交互作用都计算其对应的样本均值。

- Display Means for 框

将 Factor and Factor interactions 框中要计算均值的变量选入此框中。

- Compare main effects 复选框

当 Display Means for 框中有元素时, 该选项被激活, 用来定义是否对选中的变量进行均值的多重比较。

- Confidence interval adjustment 下拉列表

选择多重比较的方法。

- Display 复选框组

定义输出的统计量。主要包括以下统计量。

① Descriptive statistics: 描述性统计量。

② Estimates of effect size: 计算因素偏差 η^2 , 该值主要表示由该因素所导致的变异占总变异的比例。

③ Observed power: 功效检验, 用来判断试验设计的样本数是否充足, 以及部分因素是否有必要进一步研究。

④ Parameter estimates: 将各因素水平转化为哑变量之后估计其多元线型模型的系数。

⑤ Contrast coefficient matrix: 对照系数矩阵。

⑥ Homogeneity tests: 水平间的方差齐次性检验。


⑦ Spread vs. level plot: 绘制单元格的均值对应标准差、方差的散点图。

⑧ Residual plot: 绘制预测值、观察值及残差间的散点图。

⑨ Lack of fit: 检查当前模型是否能够合理描述自变量和因变量之间的关系。

⑩ General estimable function: 显示估计函数的通用表格。

- Significance level 框: 定义显著性水平。

 **注意** 当只有一个因素变量时, 执行【Univariate】过程就等价于执行单因素方差分析。

8.3.3 引例及结果解释

下面通过一个例子来介绍【Univariate】过程的操作及结果解释。

例 8.2 为考察合成纤维中收缩率与总拉伸倍数对纤维弹性有无影响, 对收缩率因子 A 取 4 个水平, 分别为 0、4、8、12。总拉伸倍数 B 也取 4 个水平, 分别为 460、520、580、640。在每个搭配 (A_i, B_j) 下做两次试验, 数据如表 8-14 所示。试分析因素 A、B 及其交互作用对弹性是否有显著影响。(数据来源:《常用统计方法》, 华东师范大学出版社)

表 8-14 纤维弹性数据表

因素 A 因素 B	A ₁	A ₂	A ₃	A ₄
B ₁	1, 3	3, 5	6, 3	5, 3
B ₂	2, 3	6, 4	9, 7	3, 2
B ₃	5, 3	8, 7	4, 5	0, 1
B ₄	7, 5	4, 4	4, 3	-1, -1

这是一个有交互作用的两因素方差分析问题，要判断因素 A、B 及其交互作用对纤维弹性的影响是否显著，执行以下操作。

STEP 01 建立如图 8-14 所示的数据文件“xianweixishu.sav”。其中“factorA”代表纤维的收缩率因子，“factorB”代表纤维的拉伸倍数，“result”代表纤维弹性。因素 A、B 的 4 个水平分别用 1、2、3、4 表示。

	factorA	factorB	result
1	1	1	1.00
2	1	1	3.00
3	1	2	3.00
4	1	2	5.00

图 8-14 数据文件“xianweixishu.sav”的数据结构

STEP 02 有交互作用的两因素方差分析，执行以下操作：

执行【Analyze】/【General Linear Model】/【Univariate】命令，弹出【Univariate】对话框	
【Dependent Variable】：result	定义纤维弹性为指标变量
【Fixed Factor】：factorA、factorB	定义纤维收缩率和拉伸倍数为因素变量
单击【Plots】按钮	弹出【Plots】对话框
【Plots】对话框：	
【Horizontal Axis】：factorA	选择纤维收缩率作为均值曲线的横坐标
【Separate Lines】：factorB	选择纤维拉伸倍数作为曲线的分组变量
单击【Add】按钮	添加定义的图形
单击【Continue】按钮	【Plots】对话框定义完成
单击【Options】按钮	弹出【Options】对话框
【Options】对话框：	
“Display Means for”框：factorA	定义估计因素 A 的均值。
单击【Continue】按钮	【Options】对话框定义完成
单击【OK】按钮	定义完成

执行以上操作之后，生成表 8-15～表 8-17。
表 8-15 给出了因素在各水平下的样本个数。从表中数据可知，纤维的收缩率和总拉伸倍数各有 4 个水平，每个水平下有 8 个样本。

表 8-15 样本个数统计表

Between-Subjects Factors		
		N
纤维收缩率	1	8
	2	8
	3	8
	4	8
纤维总 拉伸倍数	1	8
	2	8
	3	8
	4	8

表 8-16 所示是两因素方差分析表。在表格左上方给出了指标变量是“纤维弹性”。表中各列依次代表了方差来源（Source）、III型的平方和（Type III Sum of Squares）、自由度（df）、均方和（Mean Square）、F 值和显著性指标（Sig.）。表中第一行 Corrected Model 代表对方差分析模型的检验。其 Sig.取值为 0，小于 0.05，说明模型是适用的。第二行代表截距，此项可以忽略。第三、第四行分别代表因素 A、B 对指标的影响。其中因素 A 的 Sig.值大于 0.05，说明其对指标的影响不显著。因素 B 的 Sig.值为 0，小于 0.01，说明其对指标的影响高度显著。第五行代表了因素 A、B 的交互作用对指标的影响，其 Sig.值小于 0.01，说明 A、B 交互作用对指标的影响是高度显著的。

表 8-16 两因素方差分析表

Tests of Between-Subjects Effects

Dependent Variable: 纤维弹性

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	158.719 ^a	15	10.581	7.874	.000
Intercept	472.781	1	472.781	351.837	.000
factor A	8.594	3	2.865	2.132	.136
factor B	70.594	3	23.531	17.512	.000
factor A*factor B	79.531	9	8.837	6.576	.001
Error	21.500	16	1.344		
Total	653.000	32			
Corrected Total	180.219	31			

a. R Squared =.881 (Adjusted R Squared =.769)

表 8-17 所示是代表纤维收缩率（factorA）在各水平下的均值、标准误均值及 95%的置信区间。

表 8-17 纤维收缩率的均值

纤维收缩率

Dependent Variable: 纤维弹性

纤维收缩率	Mean	Std. Error	95% Confidence Interval	
			Lower Bound	Upper Bound
1	3.625	.410	2.756	4.494
2	4.500	.410	3.631	5.369
3	4.125	.410	3.256	4.994
4	3.125	.410	2.256	3.994

图 8-15 所示是两因素交互影响的均值图。均值图的横坐标代表纤维收缩率水平，纵坐标代表纤维弹性均值，且按照纤维的拉伸倍数绘制不同的折线。从图形上看，若这些折线是相交的，则认为两因素交互作用显著。若折线近似平行，则认为两因素的交互作用不显著。

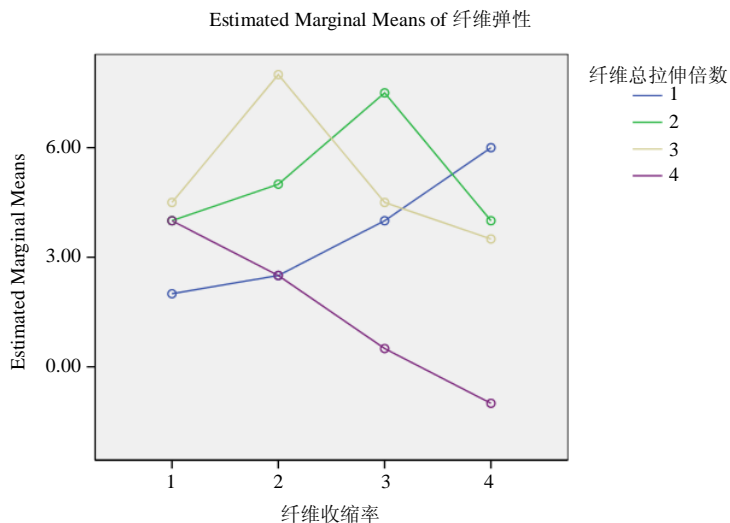


图 8-15 两因素交互影响的均值图

需要注意的是，本例由于样本个数较少，所以没有进行方差齐次性检验。从 8.1 节方差分析的适用条件可知，对于一般的方差分析问题，齐次性检验是必要的。

8.4 协方差分析——Univariate过程（2）

上一节介绍了【Univariate】过程在多因素方差检验中的应用，本节将通过具体例子介绍其在协方差分析中的应用。

8.4.1 协方差分析简介

在进行方差分析的时候，除了要分析因素变量外，其他的因素条件都要求一致或者尽

可能地接近。然而在实际问题中，这一点往往是难以控制的。譬如例 8.1 研究不同品种油菜的产量问题，试验的时候就必须要要求油菜在各块地上的种植密度完全一样。在实际问题中这一点往往难以控制。但是经验显示，种植密度的确会对产量产生影响。此时将油菜品种称为处理因素，油菜的种植密度称为混杂因素。在有混杂因素的情况下研究处理因素对指标的影响是否显著就有必要使用协方差分析的办法。

协方差分析是将方差分析和回归分析结合起来的一种统计方法。它通过回归分析来剔除其他混杂因素对指标的影响，再通过方差分析来研究处理因素对指标影响的显著性。在协方差分析中，这些混杂因素又被称为协变量。

进行协方差分析时，除了要求满足前面提到的方差分析的基本条件，还要求协变量是连续型的数值变量，且多个协变量之间要相互独立并且与因素没有交互影响。

在 SPSS 中，协方差分析也是由【Univariate】过程来完成的。对于【Univariate】过程的操作界面在上一节已经做了详细介绍，这里不再重复。

8.4.2 引例及结果解释

下面通过例子来介绍通过【Univariate】过程实现协方差分析的方法。

例 8.3 研究杨树一年生长量与施用氮肥和钾肥的关系。为了研究这种关系，一共进行了 18 个样地的栽培实验，测定杨树苗的一年生长量、初始高度、全部实验条件及实验结果，如表 8-18 所示。

表 8-18 杨树栽培实验结果

样地 序号	氮肥量	钾肥量	树苗初 始高度	生长量	样地 序号	氮肥量	钾肥量	树苗初 始高度	生长量
1	少	0	4.5	1.85	10	多	0	6.5	2.15
2	少	0	6.0	2.00	11	多	0	6.0	1.99
3	少	0	4.0	1.60	12	多	0	6.5	2.06
4	少	12.5	6.5	2.00	13	多	12.5	4.0	1.93
5	少	12.5	7.0	2.04	14	多	12.5	6.0	2.1
6	少	12.5	5.0	1.91	15	多	12.5	5.5	2.15
7	少	25	7.0	2.40	16	多	25	5.0	2.20
8	少	25	5.0	2.25	17	多	25	6.0	2.30
9	少	25	5.0	2.10	18	多	25	5.5	2.25

试根据实验数据检验氮肥量、钾肥量及树苗初始高度中哪些因素对杨树的生长有显著影响。（数据来源：《生物数学模型的统计学基础》，科学出版社）

这是一个两因素协方差分析问题，下面介绍通过【Univariate】过程来实现的步骤。

STEP 01 建立如图 8-16 所示的数据文件“yangshu.sav”。其中“N”代表氮肥量，“K”代表钾肥量，“height”代表树苗的初始高度，“grow”代表树苗的生长量。

	num	N	K	height	grow
1	1 少		0.00	4.50	1.85
2	2 少		0.00	6.00	2.00
3	3 少		0.00	4.00	1.60
4	4 少		12.50	6.50	2.00

图 8-16 数据文件“yangshu.sav”的数据结构

STEP 02 判断此问题是否可以用协方差分析的方法来处理。首先通过做散点图判断树苗初始高度与树苗生长量之间是否有线性趋势且在不同的因素水平下直线的斜率是否近似。如果各组的斜率近似则可以认为协变量和因素之间不存在交互作用。散点图如图 8-17 所示。

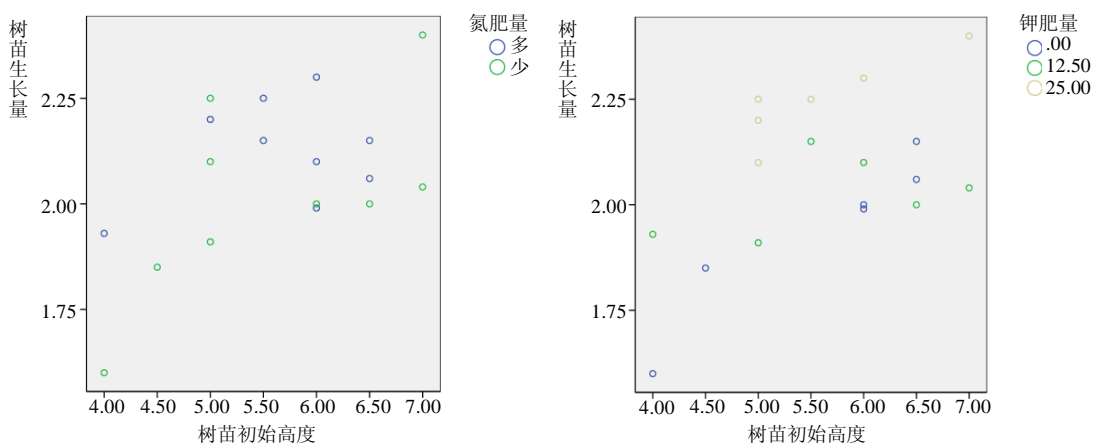


图 8-17 分别按氮肥和钾肥分组的散点图

从散点图上看，在氮肥和钾肥施肥量的各个水平上，树苗的初始高度与树苗的生长量之间都有一定的线性趋势，斜率也可以看做近似相等。但是这只是一种主观的判断，有必要通过假设检验的方法来进一步确定此问题是否符合协方差分析的条件。

STEP 03 检验不同水平下各组的斜率是否相等，以及各组方差是否齐次。执行以下操作。

执行【Analyze】/【General Linear Model】/【Univariate】命令，弹出【Univariate】对话框

【Dependent Variable】: grow

定义树苗生长量为指标变量

【Fixed Factor】: N、K

定义氮肥和钾肥为因素变量

【Covariates】: height

定义树苗初始高度为协变量

单击【Model】按钮

弹出【Model】对话框

【Model】对话框：

选中“Custom”单选框

自定义方差分析模型

【Model】: N、K、height、N*height、K*height

将 N*height、K*height 交互作用强行纳入模型

单击【Continue】按钮

【Model】对话框定义完成

单击【Options】按钮

弹出【Options】对话框

【Options】对话框：

选中“Homogeneity tests”复选框 进行方差齐次性检验

单击【Continue】按钮

【Options】对话框定义完成

单击【OK】按钮

定义完成

执行以上操作之后生成表 8-19 和表 8-20。

表 8-19 是方差齐次性检验结果。由于其 Sig. 取值大于 0.05，因此认为各组的方差齐次。

表 8-19 方差齐次性检验结果

Levene's Test of Equality of Error Variances^a

Dependent Variable: 树苗生长量

F	df1	df2	Sig.
1.018	5	12	.449

Tests the null hypothesis that the error variance of the dependent variable is equal across groups.

a. Design: Intercept + N + K + height + N*height + K*height

表 8-20 的目的是检验不同水平下各组的斜率是否相等。它主要是通过观察协变量“height”与因素变量“N”、“K”之间的交互作用是否有统计意义来实现的。在表 8-20 中，由于“height”与“N”、“K”的交互作用项 Sig. 取值分别为 0.829 和 0.226，因此认为它们之间是没有交互作用的。

表 8-20 协变量与因素变量交互作用检验

Tests of Between-Subjects Effects

Dependent Variable: 树苗生长量

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	.536 ^a	7	.077	14.356	.000
Intercept	.732	1	.732	137.070	.000
N	3.75E-006	1	3.75E-006	.001	.979
K	.025	2	.012	2.327	.148
height	.130	1	.130	24.292	.001
N*height	.000	1	.000	.049	.829
K*height	.018	2	.009	1.731	.226
Error	.053	10	.005		
Total	77.801	18			
Corrected Total	.590	17			

a. R Squared = .909 (Adjusted R Squared = .846)

通过 step 2、step 3 的检验可知，例 8.3 是满足协方差分析中关于方差齐次和协变量与因素之间没有交互作用这两个基本条件的。因此，可以用协方差分析的方法来处理例 8.3。

STEP 04 协方差分析。执行以下操作。

执行【Analyze】/【General Linear Model】/【Univariate】命令，弹出【Univariate】对话框

【Dependent Variable】: grow

定义树苗生长量为指标变量

【Fixed Factor】: N、K

定义氮肥和钾肥为因素变量

【Covariates】: height

定义树苗初始高度为协变量

单击【Model】按钮

弹出【Model】对话框

【Model】对话框：

选中“Custom”单选框

自定义方差分析模型

【Model】: N、K、height、N*K

选择分析的效应

单击【Continue】按钮

【Model】对话框定义完成

单击【Options】按钮

弹出【Options】对话框

【Options】对话框：

“Display Means for”框：N、K

定义估计因素 N、K 各水平下的修正均值

选中“Compare main effects”复选框

单击【Continue】按钮

【Options】对话框定义完成

单击【OK】按钮

定义完成

执行以上操作之后生成表 8-21～表 8-24，解释如下。

表 8-21 所示是最终的协方差分析结果。从表中可知，“N”、“K”、“height”所对应的 Sig.取值均小于 0.05。这说明氮肥、钾肥，以及树苗的初始高度都对树苗的生长量有着显著影响。但是“N*K”的 Sig.取值大于 0.05，说明氮肥和钾肥的交互作用对树苗的生长影响不明显。

表 8-21 协方差分析结果

Tests of Between-Subjects Effects

Dependent Variable: 树苗生长量

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	.538 ^a	6	.090	19.247	.000
Intercept	.627	1	.627	134.473	.000
N	.041	1	.041	8.877	.013
K	.313	2	.157	33.579	.000
N*K	.021	2	.011	2.262	.150
height	.129	1	.129	27.602	.000
Error	.051	11	.005		
Total	77.801	18			
Corrected Total	.590	17			

a. R squared=.913 (Adjusted R Squared=.866)

表 8-22 是钾肥量 3 个水平下的修正均值及其置信区间。可见钾肥的施肥量越多树苗生长得越快。表格下方的提示说明该修正均值是按照树苗初始高度-5.6111 计算得来的。

表 8-22 钾肥的修正均值

Estimates

Dependent Variable: 树苗生长量

钾 肥 量	Mean	Std.Error	95% Confidence Interval	
			Lower Bound	Upper Bound
.00	1.945 ^a	.028	1.883	2.006
12.50	2.015 ^a	.028	1.954	2.077
25.00	2.253 ^a	.028	2.192	2.315

a. Covariates appearing in the model are evaluated at the following values: 树苗初始高度=5.6111.

表 8-23 所示是钾肥组间均值两两比较的结果。通过分析表格可以发现，当钾肥的施肥量为 12.5 的时候，树苗生长量的均值与不施肥时没有显著差别。但是当施肥量达到 25 时，树苗生长量就有显著差别了。

表 8-23 钾肥修正均值两两比较

Pairwise Comparisons

Dependent Variable: 树苗生长量

(I)钾肥量	(J)钾肥量	Mean Difference (I-J)	Std. Error	Sig. ^a	95% Confidence Interval for Difference ^a	
					Lower Bound	Upper Bound
.00	12.50	-.070	.039	.102	-.157	.016
	25.00	-.308*	.039	.000	-.395	-.222
12.50	.00	.070	.039	.102	-.016	.157
	25.00	-.238*	.039	.000	-.325	-.151
25.00	.00	.308*	.039	.000	.222	.395
	12.50	.238*	.039	.000	.151	.325

Based on estimated marginal means

*. The mean difference is significant at the .05 level.

a. Adjustment for multiple comparisons: Least Significant Difference (equivalent to no adjustments).

表 8-24 所示是关于钾肥修正均值的 F 检验。其 Sig. 值小于 0.01，说明钾肥的施用量对树苗生长量的影响是非常显著的。这也与表 8-23 的结论相一致。

表 8-24 单因素检验

Univariate Tests

Dependent Variable: 树苗生长量

	Sum of Squares	df	Mean Square	F	Sig.
Contrast	.313	2	.157	33.579	.000
Error	.051	11	.005		

The F tests the effect of 钾肥量. This test is based on the linearly independent pairwise comparisons among the estimated marginal means.

对于氮肥的修正均值也由类似于表 8-22~表 8-24 的三张表得出，从分析表格可知，氮肥对树苗生长量的影响也是显著的。

8.4.3 小结

很多初学者在做协方差分析时，往往直接开始正式分析，而忽略了对协方差分析条件的检验。因此，这里必须强调在进行协方差分析之前，一定要对模型是否满足协方差分析的基本条件做出检验。同时，还要注意一点就是在正式进行协方差分析的时候，一定不能将协变量和因素变量的交互项纳入分析模型，否则可能产生完全相反的结论。

8.5 本章小结

本章介绍了方差分析在 SPSS 中的实现，主要包括单因素方差分析、多因素方差分析和协方差分析。这些方差分析方法都有各自的适用条件。所以读者在应用这些方法之前必须要检验这些方法的适用条件是否满足。这是本书反复强调的一个重点，也是初学者最容易犯的错误。

同时，请读者自行研究学习多元方差分析【Multivariate】过程。

第 9 章 相关分析

统计研究的主要内容可以分为两部分，第一部分为总体均值的差异比较，第二部分为变量间的统计关系的研究。前面所讲的 t 检验与方差分析属于第一部分。而本章要讲的相关分析则是属于第二部分。本章将通过例子学习相关分析及其在 SPSS 中的实现。本章内容包括：

- 相关分析简介
- 两变量相关分析——Bivariate 过程
- 偏相关分析——Partial 过程
- 距离分析——Distances 过程

9.1 相关分析简介

本节将概括性地介绍相关分析的基本概念、一般步骤和类型。对于各类具体的相关分析的方法，将在本章后续章节中给出详细地介绍。

9.1.1 相关分析的概念

在实际工作中，常常需要研究两个及两个以上变量的关系。比如，医学统计中研究青少年年龄和身高的关系、经济学中研究利率与股票价格的关系、农业上研究施肥量与农作物生长水平之间的关系，等等。研究这些关系主要通过相关分析和回归分析的方法来完成。

变量和变量之间的关系可以分为确定性关系和不确定性关系。所谓确定性关系是指变量之间的关系可以用精确的函数描述出来。而不确定性关系是已知变量之间存在某种联系，但是这种联系无法精确的函数描述出来。比如前面所提到的利率与股票价格的关系。一般来说，利率上调，则股票价格会下跌；利率下调，则股票价格会上升。我们仅知道在证券市场中存在这种趋势，但是股票价格与利率之间的函数关系却是很难刻画的。

如果仅仅研究变量之间相互关系的密切程度和变化趋势，并用适当的统计指标描述，这就是相关分析。如果要把变量间相互关系用函数表达出来，用一个或多个变量的取值来估计另一个变量的取值，这就是回归分析。可见相关分析是研究变量间不确定性关系的一种统计方法，而回归分析更倾向于研究变量间的确定性关系。在本章将主要介绍相关分析在 SPSS 中的实现。回归分析则在下一章介绍。

9.1.2 Correlate子菜单概述

在 SPSS 中, 相关分析主要通过【Analyze】菜单下的【Correlate】子菜单实现, 如图 9-1 所示。



图 9-1 【Correlate】子菜单

【Correlate】子菜单主要包括以下几个过程。

- ① **Bivariate:** 两变量相关分析。包括两个连续性变量之间的相关和两个等级变量之间的相关。
- ② **Partial:** 偏相关分析。当两变量的取值受其他变量的影响时, 则采用偏相关分析的方法控制其他变量的影响, 研究两变量间的相关关系。
- ③ **Distances:** 距离分析。主要用于分析同一变量内观测值之间或者多个变量之间的相似或不相似程度。

9.2 两变量相关分析——Bivariate过程

两变量相关分析在实际问题中有着广泛的应用, 通常所说的相关分析是指两变量相关分析。本节将通过例子详细介绍两变量相关分析的一般步骤、操作界面和结果解释。

9.2.1 两变量相关分析简介

两样本相关分析, 即研究两个变量之间相关关系的统计方法。两个随机变量之间的相关性是通过相关系数 ρ 来度量的。设有二维随机变量 (X, Y) , 则其相关系数由式(9.1)确定。

$$\rho_{XY} = \frac{Cov(X, Y)}{\sqrt{DX} \sqrt{DY}} = \frac{E[(X - EX)(Y - EY)]}{\sqrt{DX} \sqrt{DY}} \quad (9.1)$$

通过计算可知, $-1 \leq \rho_{XY} \leq 1$ 。若 $\rho_{XY} > 0$, 则表示 X 与 Y 正相关, 若 $\rho_{XY} < 0$, 则表示 X 与 Y 负相关。 $\rho_{XY} = 1$, 则表示 X 与 Y 正线性相关。 $\rho_{XY} = -1$, 则表示 X 与 Y 负线性相关。 $\rho_{XY} = 0$, 则表示 X 与 Y 不相关。

通常情况下, ρ_{XY} 是未知的, 而是用其样本相关系数 r 来代替 ρ_{XY} 。根据变量类型的不同, 主要有三类样本相关系数。设随机变量 (X, Y) 的 n 对样本 (x_i, y_i) ($i = 1 \cdots n$), 则三类相关系数分别表示如下。

1. Pearson相关系数

用于对定距变量的数据进行计算, 即分析两个连续性数据之间的关系。其计算公式如下:

$$r = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2 \sum_{i=1}^n (y_i - \bar{y})^2}} \quad (9.2)$$

(9.2) 式只是代表了样本的相关系数。虽然样本相关系数 r 可以作为相关系数 ρ_{XY} 的估计值,但是由于抽样误差的存在,从相关系数 $\rho_{XY} = 0$ 的总体中抽样出的样本相关系数 r 不一定为 0。此时有必要进行假设检验以确定不等于 0 的 r 是来自 $\rho_{XY} = 0$ 的总体还是 $\rho_{XY} \neq 0$ 的总体。

设 $H_0: \rho_{XY} = 0$; $H_1: \rho_{XY} \neq 0$ 。采用 t 检验的方法,在零假设下:

$$t = \frac{r}{\sqrt{\frac{1-r^2}{n-2}}} \sim t(n-2)$$

给定显著性水平 α ,若进行单侧检验,则当 $|t| > t_{\alpha, n-2}$ 时拒绝 H_0 ;若进行双侧检验,则当 $|t| > t_{\alpha/2, n-2}$ 时拒绝 H_0 ;认为 X 、 Y 是相关的。

2. Spearman秩相关系数

用于描述分类或等级变量之间、分类或等级变量与连续变量之间的相互关系。其计算公式如下:

$$r_s = 1 - \frac{6 \sum_{i=1}^n (R_i - Q_i)^2}{n(n^2 - 1)} \quad (9.3)$$

其中 R_i 表示 x_i 在 (x_1, x_2, \dots, x_n) 中的秩,所谓秩是指 x_i 在 (x_1, x_2, \dots, x_n) 中按照一定准则的排序顺序。 Q_i 表示 y_i 在 (y_1, y_2, \dots, y_n) 中的秩。

依然通过 t 检验的方法来确定不等于 0 的 r_s 是来自 $\rho_{XY} = 0$ 的总体还是 $\rho_{XY} \neq 0$ 的总体。设 $H_0: \rho_{XY} = 0$; $H_1: \rho_{XY} \neq 0$ 。此时的检验统计量 t 表示如下:

$$t = r_s \sqrt{\frac{n-2}{1-r_s^2}} \sim t(n-2)$$

给定显著性水平 α ,若进行单侧检验,则当 $|t| > t_{\alpha, n-2}$ 时拒绝 H_0 ;若进行双侧检验,则当 $|t| > t_{\alpha/2, n-2}$ 时拒绝 H_0 ;认为 X 、 Y 是相关的。

研究表明,在正态分布假定下, Spearman 秩相关系数与 Pearson 相关系数在效率上是等价的,而对于非正态分布或者分布不明的数据,则采用 Spearman 秩相关系数更合适。

3. Kendall相关系数

Kendall 相关系数与 Spearman 秩相关系数类似,也是用于描述分类或等级变量之间、分类或等级变量与连续变量之间的相互关系。这种方法是通过两变量 (x_i, y_i) ($i=1 \cdots n$) 是否协同一致来检验两变量之间是否存在相关性。 (x_i, y_i) 与 (x_j, y_j) 协同即是指 $(x_i - x_j)(y_i - y_j) > 0$ 。构造如下的 Kendall τ 检验统计量:

$$\tau = \frac{2}{n(n-2)} \sum_{1 \leq i < j \leq n} \text{sign}((x_i - x_j)(y_i - y_j)) \quad (9.4)$$

通过计算可知 $-1 \leq \tau \leq 1$ 。设 $H_0: \rho_{XY} = 0$; $H_1: \rho_{XY} \neq 0$ 。给定显著性水平 α ，查 Kendall 表：若进行单侧检验，则当 $|\tau| > \tau_{\alpha, n}$ 时拒绝 H_0 ；若进行双侧检验，则当 $|\tau| > \tau_{\alpha/2, n}$ 时拒绝 H_0 ；认为 X 、 Y 是相关的。

9.2.2 Bivariate过程的操作界面

首先介绍【Bivariate】过程的操作界面。执行【Analyze】/【Correlate】/【Bivariate】命令，弹出如图 9-2 所示的【Bivariate】对话框。

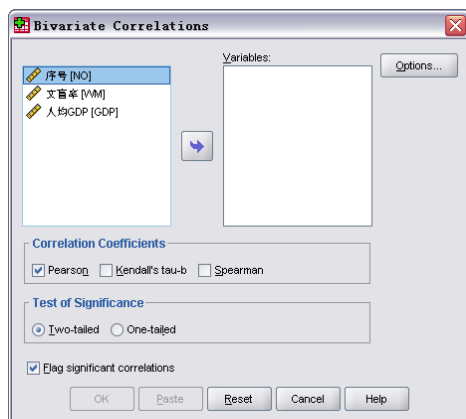


图 9-2 【Bivariate】对话框

该对话框主要由以下几部分组成。

1. 候选变量框

即左侧变量列表框。

2. 【Variables】

选择要进行相关分析的变量。至少选入两个变量。如果选入变量个数大于两个，则对其分别进行两两相关分析。

3. 【Correlation Coefficients】复选框组

选择要计算的相关系数。

- ① Pearson 复选框：对于连续性变量计算 Pearson 相关系数，系统默认选中此项。
- ② Kendall's tau-b 复选框：对于分类或等级变量计算 Kendall 相关系数。
- ③ Spearman 复选框：对于分类或等级变量计算 Spearman 相关系数。

4. 【Test of Significance】单选框组

定义相关系数的检验方法。包括双侧检验（Two-tailed）和单侧检验（One-tailed）两种方法，具体选择哪种检验方法要依赖于实际问题。系统默认进行双侧检验。

5. “Flag significant correlations” 复选框

用“*”标记有统计学意义的相关系数。如果 $\text{Sig.} < 0.05$ 则用一个“*”标记，若 $\text{Sig.} < 0.01$ 则用两个“*”标记。

6. 【Options】

单击【Options】按钮，弹出如图 9-3 所示的【Options】对话框。

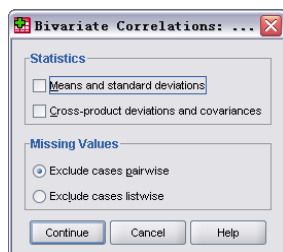


图 9-3 【Options】对话框

该对话框主要用于选择输出统计量和定义缺失值的处理方式，包括以下两部分。

• Statistics 复选框组

给出可以选择输出的统计量。

① Means and standard deviations 复选框：输出各个变量的样本均值及标准差。

② Cross-product deviations and covariances 复选框：输出各对变量的交叉积及协方差矩阵。

• Missing Values 单选框组

定义缺失值处理方式。

① Exclude cases pairwise 单选框：仅当数据中要分析的变量值缺失时才剔除该数据，系统默认选中此项。

② Exclude cases listwise 单选框：只要数据中有变量值缺失就剔除该数据。

9.2.3 引例及结果解释

下面通过一个例子介绍【Bivariate】过程的操作及结果解释。

例 9.1 从中国 30 个省区抽样的文盲率（单位：1%）和各省人均 GDP（单位：元）的数据如表 9-1 所示。

表 9-1 抽样数据表

序 号	文 盲 率	人 均 GDP	序 号	文 盲 率	人 均 GDP
1	7.33	15044	16	10.94	6468
2	10.80	12270	17	20.97	3881
3	15.60	5345	18	16.40	3715
4	8.86	7730	19	16.59	4032
5	9.70	22275	20	17.40	5122

续表

序 号	文 盲 率	人 均 GDP	序 号	文 盲 率	人 均 GDP
6	18.52	8447	21	14.12	4130
7	17.71	9455	22	18.99	3763
8	21.24	8136	23	30.18	2093
9	23.20	6834	24	28.48	3715
10	14.24	9513	25	61.13	2732
11	13.82	4081	26	21.00	3313
12	17.97	5500	27	32.88	2901
13	10.00	5163	28	42.14	3748
14	10.15	4220	29	25.02	3731
15	17.05	4259	30	14.65	5167

问文盲率与人均 GDP 之间是否相关？是正相关还是负相关？（数据来源：《非参数统计》，中国人民大学出版社）

对于例 9.1 执行以下操作。

STEP 01 建立如图 9-4 所示的数据文件“WM&GDP.sav”。其中“WM”代表文盲率，“GDP”代表人均 GDP。

	NO	WM	GDP
1	1	7.33	15044.00
2	2	10.80	12270.00
3	3	15.60	5345.00
4	4	8.86	7730.00

图 9-4 数据文件“WM&GDP.sav”的数据结构

STEP 02 作出文盲率与 GDP 之间的散点图，初步判断两变量是否有相关关系及该关系是否呈线性。

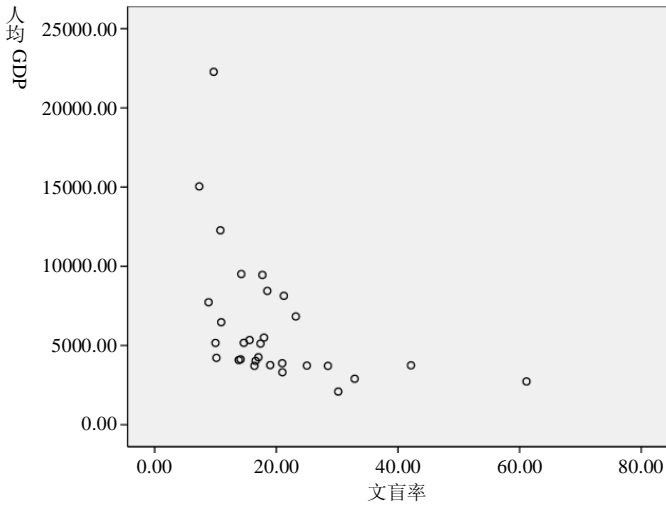


图 9-5 文盲率与人均 GDP 的散点图

从图形上看，二者的确存在线性相关关系。只有线性相关的关系确定后才能继续进行下一步分析。因此，在进行相关分析之前的预分析过程也是十分重要的。而初学者往往容易忽略这一点。

STEP 03 两因素的相关分析。执行以下操作：

执行【Analyze】/【Correlate】/【Bivariate】命令，弹出【Bivariate】对话框

【Variables】：WM、GDP 选择分析文盲率与人均 GDP 之间相关关系

选中“Pearson”和“Spearman”复选框 计算两类相关系数

选中“Two-tailed”单选框 对相关系数进行双侧检验

选中“Flag significant correlations”选项 标识有统计意义的相关系数

单击【Options】按钮 弹出【Options】对话框

【Options】对话框：

选中“Means and standard deviations”复选框 输出变量的均值和标准差

单击【Continue】按钮 【Options】对话框定义完成

单击【OK】按钮 定义完成

执行以上操作之后，生成表 9-2～表 9-4 所示数据，分别解释如下。

表 9-2 给出了两变量的描述性统计量。从左至右分别为均值（Mean）、标准差（Std.Deviation）和样本数（N）。

表 9-2 描述性统计量

Descriptive Statistics			
	Mean	Std. Deviation	N
文盲率	19.5693	11.01357	30
人均 GDP	6226.1000	4219.02924	30

表 9-3 所示是 Pearson 相关系数及其显著性检验结果。由于其相关系数为-0.449，相关系数的 Sig. 为 0.013，大于 0.01 小于 0.05。所以相关系数用“*”标记，说明文盲率与人均 GDP 的相关性是显著的。

表 9-3 Pearson 相关系数

Correlations			
		文盲率	人均 GDP
文盲率	Pearson Correlation	1	-.449*
	Sig. (2-tailed)		.013
	N	30	30
人均 GDP	Pearson Correlation	-.449*	1
	Sig. (2-tailed)	.013	
	N	30	30

*. Correlation is significant at the 0.05 level (2-tailed).

表 9-4 所示是 Spearman 相关系数及其显著性检验的结果。其相关系数为-0.631，相关系数的 Sig. 为 0.000，小于 0.01。所以相关系数用“***”标记，说明文盲率与人均 GDP 的

相关性是高度显著的。

表 9-4 Spearman 相关系数

Correlations			文 盲 率	人 均 GDP
Spearman's rho	文盲率	Correlation Coefficient	1.000	-.631**
		Sig. (2-tailed)	.	.000
		N	30	30
	人均 GDP	Correlation Coefficient	-.631**	1.000
		Sig. (2-tailed)	.000	.
		N	30	30

** . Correlation is significant at the 0.01 level (2-tailed).

从表 9-3 和表 9-4 的分析结果可知，无论是用 Pearson 相关系数还是用 Spearman 相关系数，都能得出文盲率与人均 GDP 是负相关的结论。

通过例 9.1 的学习，可以发现两样本相关分析的一个特点，即进行分析的两变量间是处于一个平等的地位。X 与 Y 的相关关系同 Y 与 X 的相关关系是完全一致的。这一点对所有的相关分析都适用，也是相关分析和回归分析的重要区别之一。

9.3 偏相关分析——Partial过程

9.2 节介绍了两样本相关分析及其在 SPSS 中的实现。然而由于客观世界的复杂性，分析两样本相关关系的时候可能还会受到其他因素的影响。因此，在实际问题中又引入了偏相关分析的方法。本节将通过例子介绍偏相关分析的概念、操作界面及结果解释。

9.3.1 偏相关分析简介

在实际问题中，两变量间的相关关系往往还要受到其他因素的影响。这些影响有时候会使相关分析的结果变得不那么可靠。比如研究冰川地区河水的径流量与降雨量的关系。对于冰川地区的河水径流量，除了受到降雨量的影响之外，温度也是一个重要因素。如果不考虑温度影响，那么最后分析的结果可能与实际情况不符。因此，又引入了偏相关分析的方法。所谓偏相关分析，是指在研究两变量之间的相关关系时，将与这两个变量有联系的其他变量控制不变的统计方法。

根据控制变量的个数，偏相关分析分为零阶偏相关分析、一阶偏相关分析、二阶偏相关分析……。所谓零阶偏相关分析是指没有控制变量的相关分析，这就等同于一般的相关分析。一阶偏相关分析是指有一个控制变量的相关分析，二阶偏相关分析是指有两个控制变量的偏相关分析，其他高阶偏相关分析都是以此类推。

进行偏相关分析时首先也是要计算偏相关系数。偏相关系数在数值上与一般相关系数往往是不同的。在计算一般相关系数的时候，只考虑两变量之间的关系，其他的变量都不予考虑。而计算偏相关系数时，要考虑其他变量的影响，只是把其他变量当做常数。对于

偏相关系数，还是能通过 t 检验来检验其显著性。由于偏相关系数及其检验统计量的表达式比较复杂，这里不再具体介绍。有兴趣的读者可以参阅相关书籍或者查看 SPSS 的帮助文档。

9.3.2 Partial过程的操作界面

首先介绍【Partial】过程的操作界面。执行【Analyze】/【Correlate】/【Partial】命令，弹出如图 9-6 所示的【Partial】对话框。

该对话框主要由以下几部分组成。

1. 候选变量框

即左侧变量列表框。

2. 【Variables】

选择要进行偏相关分析的变量，至少选入两个变量。如果选入变量个数大于两个，则分别对其进行两两偏相关分析。

3. 【Controlling for】

选择偏相关分析中的控制变量，如果不选的话，则等同于进行一般的相关分析。

4. 【Test of Significance】单选框组

定义相关系数的检验方法。包括双侧检验（Two-tailed）和单侧检验（One-tailed）两种。系统默认为进行双侧检验。

5. “Display actual significance level” 选项

选择是否给出真实的显著性水平值，系统默认选中此项。

• 【Options】

单击【Options】按钮，弹出如图 9-7 所示的【Options】对话框。

该对话框主要用于选择输出统计量和定义缺失值的处理方式，包括以下两部分。

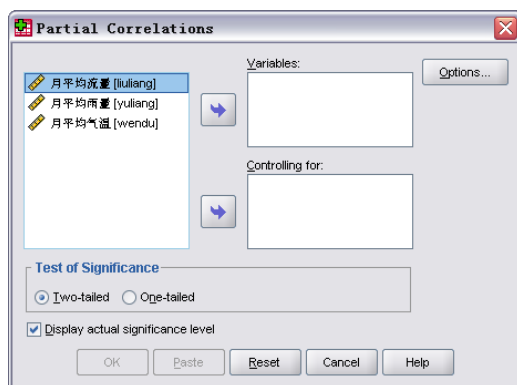


图 9-6 【Partial】对话框

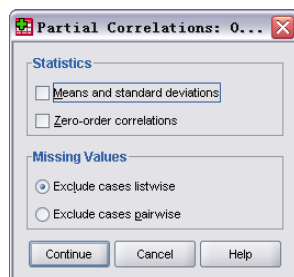


图 9-7 【Options】对话框

• Statistics 复选框组

给出可以选择输出的统计量。

① Means and standard deviations: 输出各变量的样本均值和标准差。

② Zero-order correlations: 给出包括控制变量在内的所有变量的相关矩阵。

• Missing Values 单选框组

定义缺失值的处理方式。

① Exclude cases listwise: 只要数据中有变量值缺失就剔除该数据, 系统默认选中此项。

② Exclude cases pairwise: 仅当数据要分析的变量值缺失时才剔除该数据。

9.3.3 引例及结果解释

下面通过一个例子来介绍【Partial】过程的操作及结果解释。

例9.2 已知有某河流的12个月的月平均流量观测数据和该河流所在地区当年的月平均降雨量和月平均温度观测数据, 如表9-5所示。试分析温度与河水流量之间的相关关系。

表9-5 观测数据表

月 份	月平均流量	月平均雨量	月平均气温
1 月	0.50	0.10	-8.80
2 月	0.30	0.10	-11.00
3 月	0.40	0.40	-2.40
4 月	1.40	0.40	6.90
5 月	3.30	2.70	10.60
6 月	4.70	2.40	13.90
7 月	5.90	2.50	15.40
8 月	4.70	3.00	13.50
9 月	0.90	1.30	10.00
10 月	0.60	1.80	2.70
11 月	0.50	0.60	-4.80
12 月	0.30	0.20	-6.00

河流流量除了受温度影响之外, 降雨量也是一个重要的影响因素。那么要研究流量与温度的关系, 有必要剔除降雨量的影响。这就是一个把降雨量作为控制变量的偏相关分析问题。执行以下操作。

STEP 01 建立如图9-8所示的数据文件“liuliang.sav”。其中“liuliang”代表河流的月平均流量, “yuliang”代表该地区的月平均降雨量, “wendu”代表该地区的月平均温度。

	yuefen	liuliang	yuliang	wendu
1	1月	0.50	0.10	-8.80
2	2月	0.30	0.10	-11.00
3	3月	0.40	0.40	-2.40
4	4月	1.40	0.40	6.90

图9-8 数据文件“liuliang.sav”的数据结构

STEP 02 偏相关分析。执行以下操作:

执行【Analyze】/【Correlate】/【Partial】命令，弹出【Partial】对话框	
【Variables】：liuliang、wendu	选择分析流量与温度间的偏相关关系
【Controlling for】：yuliang	选择降雨量为控制变量
选中“Two-tailed”单选框	对偏相关系数进行双侧检验
选中“Display actual significance level”选项	给出真实的显著性水平值
单击【Options】按钮	弹出【Options】对话框
【Options】对话框：	
选中“Zero-order correlations”复选框	输出各变量的相关矩阵
单击【Continue】按钮	【Options】对话框定义完成
单击【OK】按钮	定义完成

执行以上操作之后，生成表 9-6。表 9-6 可以看作是两个部分。上半部分是对各变量做两两一般的相关分析的结果。表格的注释 a 说明此时相关系数是按照 Pearson 相关系数计算的。从一般的相关分析结果来看，月平均气温与月平均流量的相关系数取值为 0.836，其 Sig. 值为 0.001，小于 0.01。说明在一般的相关分析下，月平均气温和月平均流量的相关性是高度显著的。表 9-6 的下半部分是以月平均降雨量作为控制变量的偏相关分析结果。从表格中可以看出，此时月平均气温与月平均流量的偏相关系数取值为 0.365，其 Sig. 值为 0.27，大于 0.05。说明此时月平均气温和月平均流量的相关性不显著。

表 9-6 一般相关分析和偏相关分析结果表

Correlations					
Control Variables			月平均流量	月平均气温	月平均雨量
-none. ^a	月平均流量	Correlation	1.000	.836	.855
		Significance(2-tailed)	.	.001	.000
		df	0	10	10
	月平均气温	Correlation	.836	1.000	.867
		Significance(2-tailed)	.001	.	.000
		df	10	0	10
	月平均雨量	Correlation	.855	.867	1.000
		Significance(2-tailed)	.000	.000	.
		df	10	10	0
月平均雨量 月平均流量		Correlation	1.000	.365	
		Significance(2-tailed)	.	.270	
		df	0	9	
月平均气温		Correlation	.365	1.000	
		Significance(2-tailed)	.270	.	
		df	9	0	

a. Cells contain zero-order (Pearson) correlations.

从表 9-6 最后的分析结果可以看出，当两个变量还受其他变量影响的时候，用一般的相关分析和偏相关分析得出了完全不一样的结论。可见，在处理实际问题的时候，一定要根据问题的背景，剔除一些对分析结果有影响的变量。

9.4 距离分析——Distances过程

前面介绍的两样本相关分析和偏相关分析都是研究两样本之间的相关关系。对于两个变量，除了研究它们之间的相关关系外，研究两者的近似程度也是十分重要的问题。距离分析是用于研究变量是否近似的一种相关分析方法。本节将通过例子介绍距离分析的概念、操作界面及结果解释。

9.4.1 距离分析简介

在模式识别中，为了能划分模式的类别，首先必须定义模式的相似性测度，以此来描述各模式之间特征的相似程度。距离分析是用来描述同一变量内观测值之间或者是多个变量之间的相似或不相似程度的统计方法。

在距离分析中通常用距离测度 d 来描述观测值或变量间的不相似程度，用相似测度来描述观测值或变量间的相似程度。由于距离测度是以两个矢量矢端的距离作为考虑的基础，因此距离测度值是两矢量各相应分量之差的函数。距离测度越小，说明两观测值或变量越相似。而相似测度是以两矢量的方向是否近似作为考虑的基础，此时矢量的长度就不重要了。相似测度值越大，说明两观测值或变量越相似。

SPSS 中根据变量类型的不同，提供了极为丰富的距离测度和相似测度，这些都将在下一小节给出详细介绍。

9.4.2 Distances过程的操作界面

首先介绍【Distances】过程的操作界面。执行【Analyze】/【Correlate】/【Distances】命令，弹出如图 9-9 所示的【Distances】对话框。

该对话框主要由以下几部分组成。

1. 候选变量框

即左侧变量列表框。

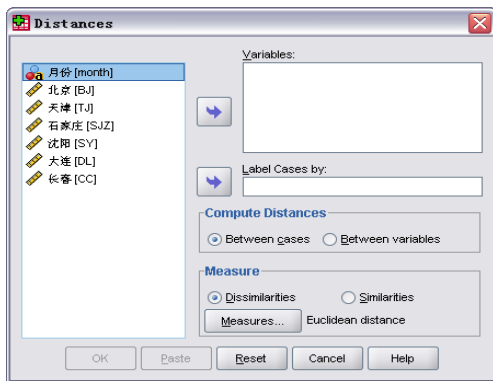


图 9-9 【Distances】对话框

2. 【Variables】

选择要进行距离分析的变量。至少选入两个变量。

3. 【Label Cases by】

选择标识变量。在最后的输出结果中，给出标识变量的取值，方便阅读。

4. 【Compute Distances】单选框组

定义距离分析的类型。

- ① Between cases 单选框：定义对观测值进行距离分析。
- ② Between variables 单选框：定义对变量进行距离分析。

5. 【Measure】单选框组

选择距离分析的测度类型。

- ① Dissimilarities 单选框：计算不相似性测度。
- ② Similarities 单选框：计算相似性测度。

6. 【Measures】

单击【Measures】按钮，弹出【Measures】对话框。该对话框主要用于具体定义距离分析的测度类型。根据【Measures】单选框组选择的不同，会弹出不同的对话框。如果选择的是 Dissimilarities 单选框，则对应的是如图 9-10 所示的【Measures】对话框；若选择的是 Similarities 单选框，则对应如图 9-11 所示的【Measures】对话框。

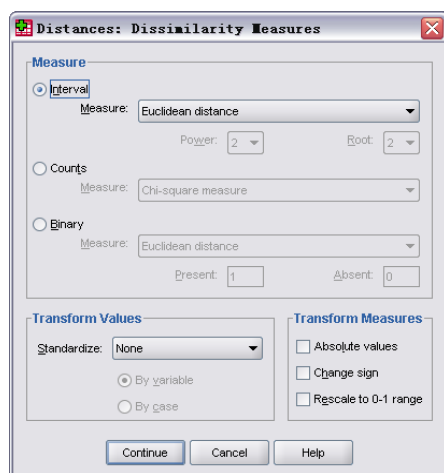


图 9-10 【Measure】对话框 1

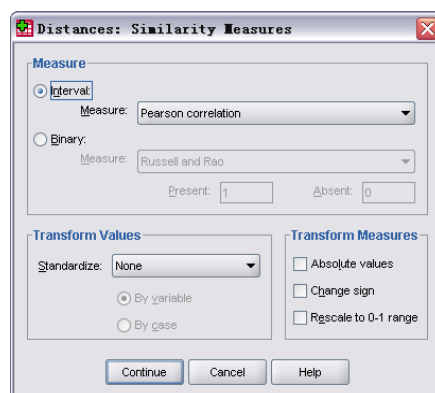


图 9-11 【Measure】对话框 2

首先介绍不相似测度下的【Measures】对话框（如图 9-10 所示）。该对话框主要由以下几部分组成。

• Measure 单选框组

根据变量或观测值数据类型的不同，选择不同的不相似测度即距离测度指标。

① Interval 单选框：计算定距变量的距离测度。选择该项，激活 Interval 下拉列表。该列表主要包括如表 9-7 所示的几个具体指标。设 $x = (x_1, x_2, \dots, x_n)'$ ， $y = (y_1, y_2, \dots, y_n)'$ 。

表 9-7 Interval 下拉列表

SPSS 中表示	名 称	公 式
Euclidean distance	欧式距离	$d(x, y) = [\sum_{i=1}^n (x_i - y_i)^2]^{1/2}$
Squared Euclidean distance	欧式距离的平方	$d(x, y) = \sum_{i=1}^n (x_i - y_i)^2$
Chebyshev	切氏距离	$d(x, y) = \max_i x_i - y_i $
Block	绝对值距离	$d(x, y) = \sum_{i=1}^n x_i - y_i $
Minkowski	明氏距离	$d(x, y) = [\sum_{i=1}^n x_i - y_i ^m]^{1/m}$ (m 为待定参数)
Customized	用户自定义距离	$d(x, y) = [\sum_{i=1}^n x_i - y_i ^p]^{1/q}$ (p 、 q 为待定参数)

② Counts 单选框：计算分类变量的距离测度。选择该项，激活 Counts 下拉列表。该列表主要包括如表 9-8 所示的两项指标。

表 9-8 Counts 下拉列表

SPSS 中表示	名 称	公 式
Chi-square measure	χ^2 测度	$d_{chi}(x, y) = \sqrt{\sum_{i=1}^n \frac{(x_i - E(x_i))^2}{E(x_i)} + \sum_{i=1}^n \frac{(y_i - E(y_i))^2}{E(y_i)}}$
Phi-square measure	ϕ^2 测度	$d_{phi}(x, y) = \frac{d_{chi}(x, y)}{\sqrt{n}}$

③ Binary 单选框：计算二元变量的距离测度。对二元变量计算距离测度的时候，首先建立如表 9-9 所示的列联表。其中“Present”表示该变量具有某项特征。“Absent”表示该变量不具有某项特征。a、b、c、d 分别表示满足条件的变量对个数。

表 9-9 二元变量列联表

2 1	Present	Absent
Present	a	b
Absent	c	d

在 SPSS 中，默认当变量取值为 1 时代表“Present”，变量取值为 0 时代表“Absent”。该取值可以通过 Binary 下拉列表下方的 Present 框和 Absent 框调整。

Binary 下拉列表主要包括如表 9-9 所示的几项指标。

表 9-10 Binary 下拉列表

SPSS 中表示	名 称	公 式
Euclidean distance	欧式距离	$d(x, y) = \sqrt{b + c}$
Squared Euclidean distance	欧式距离的平方	$d(x, y) = b + c$
Size difference	大小差测度	$d(x, y) = \frac{(b - c)^2}{(a + b + c + d)^2}$

续表

SPSS 中表示	名 称	公 式
Pattern difference	型差异测度	$d(x, y) = \frac{bc}{(a+b+c+d)^2}$
Variance	变差测度	$d(x, y) = \frac{b+c}{4(a+b+c+d)}$
Shape	形状测度	$d(x, y) = \frac{(a+b+c+d)(b+c) - (b-c)^2}{(a+b+c+d)^2}$
Lance and Williams	Lance—Williams 测度	$d(x, y) = \frac{b+c}{2a+b+c}$

• Transform Values 组

定义数据标准化方法。当数据的量不一致时会使分析结果发生偏差，此时有必要对数据进行标准化。主要包括以下标准化方法。

- ① None：不变化。
- ② Z scores：进行 Z 变换。
- ③ Range -1 to 1：将数据标准化到-1~1 之间。
- ④ Range 0 to 1：将数据标准化到 0~1 之间。
- ⑤ Maximum magnitude of 1：将数据标准化后使其最大值为 1。
- ⑥ Mean of 1：将数据标准化之后使其均值为 1。
- ⑦ Standard deviation of 1：将数据标准化之后使其标准差为 1。

当选择了要对数据标准化之后，还应选择是对变量进行标准化（By variable）还是对观测值进行标准化（By case）。

• Transform Measures 复选框组

定义对计算出来的距离测度作进一步的转化，主要包括以下几种方法。

- ① Absolute values 复选框：绝对值转换法。
- ② Change sign 复选框：变号转换法。
- ③ Rescale to 0-1 range 复选框：将距离测度转换到（0，1）区间。

相似测度下的【Measures】对话框（如图 9-10 所示）与不相似测度下的【Measures】对话框（如图 9-11 所示）在很多地方是一致的。下面仅介绍其不同的地方即 Measure 单选框组。此时 Measure 单选框组的作用是根据变量或观测值数据类型不同，选择不同的相似测度。主要有以下选项。

① Interval 单选框：计算定距变量的相似测度。选择该项，激活 Interval 下拉列表。该列表主要包括如表 9-11 所示的两项指标。设 $x = (x_1, x_2, \dots, x_n)'$ ， $y = (y_1, y_2, \dots, y_n)'$ 。

表 9-11 Interval 下拉列表

SPSS 中表示	名 称	公 式
Pearson Correlation	Pearson 相关系数	$r(x, y) = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2 \sum_{i=1}^n (y_i - \bar{y})^2}}$
Cosine	角度相似系数	$\cos(x, y) = \frac{x'y}{[(x'x)(y'y)]^{1/2}}$

② Binary 单选框：计算二元变量的相似测度。在 SPSS 中，共提供了 20 种二元变量的相似测度，这里就不再一一详述了。

9.4.3 引例及结果解释

下面通过一个例子来介绍【Distances】过程的操作及结果解释。

例 9.3 已知有我国六个城市 2004 年各月的日照时数数据如表 9-12 所示。请分析各城市日照数是否近似。（数据来源：《2005 年中国统计年鉴》，中国统计出版社）

表 9-12

城市 月份	北 京	天 津	石 家 庄	沈 阳	大 连	长 春
1	194.70	161.70	193.80	165.40	163.50	194.10
2	213.50	185.20	219.20	180.70	195.30	165.00
3	243.60	166.80	220.90	231.70	223.10	246.70
4	248.20	214.30	240.90	245.30	276.90	266.80
5	253.30	221.00	277.90	219.30	243.40	246.20
6	202.00	182.50	213.40	230.30	190.00	265.50
7	203.20	179.50	185.40	133.00	228.50	183.50
8	187.40	149.80	152.10	198.30	174.00	282.70
9	198.90	178.70	203.40	211.10	202.70	232.70
10	225.20	194.70	220.70	229.90	228.40	236.20
11	201.40	172.80	197.50	132.20	172.90	138.70
12	144.00	119.10	97.90	114.50	167.00	144.50

采用距离分析的方法分析这几个城市之间的日照时数是否近似，执行以下操作。

STEP 01 建立如图 9-12 所示的数据文件“rizhao.sav”。

	month	BJ	TJ	SJZ	SY	DL	CC
1	Jan	194.70	161.70	193.80	165.40	163.50	194.10
2	Feb	213.50	185.20	219.20	180.70	195.30	165.00
3	Mar	243.60	166.80	220.90	231.70	223.10	246.70
4	Apr	248.20	214.30	240.90	245.30	276.90	266.80

图 9-12 数据文件“rizhao.sav”的数据结构

STEP 02 距离分析，执行以下操作：

执行【Analyze】/【Correlate】/【Distances】命令，弹出【Distances】对话框

【Variables】：BJ、TJ、SJZ、SY、DL、CC

选择分析各城市的近似程度

【Compute Distances】：选中 Between variables 单选框

定义进行变量间距离分析

【Measure】单选框组：选中 Dissimilarities 单选框

定义计算不相似测度

单击【Measures】按钮

弹出【Measures】对话框

【Measures】对话框：

Interval 下拉列表：Euclidean distance	选择计算变量间欧式距离
单击【Continue】按钮	【Measures】对话框定义完成
单击【OK】按钮	定义完成

执行以上操作之后，生成表 9-13 和表 9-14。表 9-13 所示是变量的观测值数及其缺失值情况。

表 9-13 数据摘要

Case Processing Summary

Cases					
Valid		Missing		Total	
N	Percent	N	Percent	N	Percent
12	100.0%	0	.0%	12	100.0%

表 9-14 所示是距离分析的结果表。表格下方注释“This is a dissimilarity matrix”，表明此时距离分析采用的是不相似测度。表格第一行“Euclidean distance”，表明表格中的不相似测度采用的是欧氏距离。当两变量间的欧式距离越大，说明其差别越大，反之亦然。

表 9-14 距离分析结果表

Proximity Matrix

	Euclidean Distance					
	北 京	天 津	石 家 庄	沈 阳	大 连	长 春
北京	.000	122.933	71.280	122.139	70.542	146.479
天津	122.933	.000	111.350	126.363	121.427	205.540
石家庄	71.280	111.350	.000	125.332	110.928	178.273
沈阳	122.139	126.363	125.332	.000	133.006	121.829
大连	70.542	121.427	110.928	133.006	.000	157.159
长春	146.479	205.540	178.273	121.829	157.159	.000

This is a dissimilarity matrix

9.5 本章小结

本章介绍了相关分析【Correlate】子菜单，详细介绍了以下几个过程：

- Bivariate 过程，两变量相关分析；
- Partial 过程，偏相关分析；
- Distances 过程，距离分析。

其中两变量相关分析是最简单最常用的。但是当两变量的取值受其他变量的影响时，如果还采用简单的两变量相关分析的话就有可能得出错误的结论。这时则要采用偏相关分析的方法控制其他变量的影响，研究两变量间的相关关系。

相关分析和偏相关分析主要用于研究两变量间的相关关系。如果要分析同一变量内观测值之间或者多个变量之间的相似或不相似程度，则需要采用距离分析。

第 10 章 回归分析

回归分析是实际工作中应用最广泛的统计方法之一。SPSS 的【Regression】子菜单包括了所有常用回归分析方法。本章对各类常用回归分析的数学背景给出详细讲解，并且通过实际例子介绍在 SPSS 环境下各类回归分析的具体操作和结果分析。本章内容包括：

- 回归分析简介
- 线性回归——Linear 过程
- 曲线拟合——Curve Estimation 过程
- 二分类变量 Logistic 回归——Binary Logistic 过程
- 非线性回归——Nonlinear 过程

10.1 回归分析简介

本节主要概括性地介绍回归分析的概念、应用背景和分类等。由于回归分析包含的方法极其丰富且千差万别。因此，各类具体的回归分析方法将在本章的后续几节具体介绍。

10.1.1 回归分析的概念

最初回归分析的概念主要是针对两个变量。如果两个变量之间存在着较高的相关，则可以试着从一个变量的变化去推断另一个变量的变化。通常把一个变量作为自变量，把另一个变量作为因变量，建立两者的数学表达式，从自变量去估计因变量的取值，这个过程叫做回归分析。

现在回归分析的概念则包含了更多的内容，概括地讲，回归分析是描述两个或两个以上变量间关系的一种统计方法。上一章所提到的相关分析也是研究两个或两个以上变量间关系的一种统计方法。因此，在具体讲解回归分析之前，我们有必要弄清楚回归分析和相关分析的区别。

首先给出一个简单的例子来说明回归分析与相关分析的区别。已知有如表 10-1 所示的数据：

表 10-1 回归分析与相关分析的区别

X	1	2	3	4	5	6	7	8
Y1	1	2	3	4	5	6	7	8
Y2	2	4	6	8	10	12	14	16

分析表中数据可知， $(X, Y1)$ 、 $(X, Y2)$ 这两组数据的相关系数是一样的。但是，显然不能因为这两组数据的相关系数一致，就说 X 变化对 $Y1$ 的影响与 X 变化对 $Y2$ 的影响也是一致的。 X 每增加一个单位， $Y1$ 增加一个单位， $Y2$ 却增加两个单位。这一信息是相关分析无法描述的，需要用回归分析来研究。

回归分析是指通过提供变量之间的数学表达式来定量描述变量间相关关系的数学过程。这一数学表达式通常称为经验公式。我们不仅可以利用概率统计知识对这个经验公式的有效性进行判定。同时还可以利用这个经验公式，根据自变量的取值来预测因变量的取值。如果是多个因素作为自变量的时候，还可以通过进行因素分析，找出哪些自变量对因变量的影响是显著的，哪些是不显著的。

10.1.2 回归分析的应用

在实际工作中回归分析的应用范围很广。目前应用最多的是生物统计和医学统计。例如，估计各类微量元素的摄入量对人体血红蛋白含量的影响、一种新药对人体各项指标所造成的副作用有多大，等等。

但是回归分析的应用不仅仅局限在这两个领域。因为回归分析可以求出自变量与因变量之间的经验公式。所以，只要需要定量分析多变量之间相关关系时，回归分析都是必不可少的。现在所流行的数据挖掘技术，回归分析也是必不可少的。

回归分析在数据分析中主要有预测和控制两大功能。通过对已知训练数据进行回归分析得出经验公式，利用经验公式就可以在已知自变量的情况下预测因变量的取值。实际问题中往往是根据预测结果来进行控制调整。例如在商品流通领域，经常用回归分析来分析商品价格与商品需求量之间的关系，以便对商品的价格和需求量进行控制。

本章在具体介绍各类回归分析方法的时候都会给出一些实际问题的例子。这样，读者就能了解到回归分析在数据分析中的强大功能了。

10.1.3 回归分析的类型

回归分析按照经验公式的函数类型可以分为线性回归和非线性回归。若回归分析的经验公式是线形函数，则称为线性回归。若经验公式是非线性函数，则称为非线性回归。

按自变量个数可以将回归分析分为一元回归和多元回归。其中一元回归是指只有一个自变量的回归分析。有两个或两个以上自变量的回归分析称为多元回归。

按自变量和因变量的类型回归分析又可以分为一般的回归分析、含有哑变量的回归分析和 Logistic 回归分析。通常所遇到的问题，自变量和因变量都是定量变量，这种回归是一般的回归分析。若因变量是定量变量，自变量含有定性变量，则称为含有哑变量的回归分析。若因变量是定性变量的回归分析，则称为 Logistic 回归分析。

如图 10-1 所示，SPSS 的【Regression】子菜单几乎囊括了所有的回归分析方法。

【Regression】子菜单各过程的具体作用如下。

- ① **Linear**：线性回归。既可以进行一元线性回归，也可以进行多元线性回归，是【Regression】子菜单最常用的过程。
- ② **Curve Estimation**：曲线拟合。

- ③ Partial Least Squares: 偏最小二乘法。
- ④ Binary Logistic: 因变量为二分类变量的 Logistic 回归。
- ⑤ Multinomial: 因变量为多分类无序变量的 Logistic 回归。
- ⑥ Ordinal: 因变量为多分类有序变量的 Logistic 回归。
- ⑦ Probit: 概率单元回归。
- ⑧ Nonlinear: 非线性回归。
- ⑨ Weight Estimation: 加权最小二乘回归。
- ⑩ 2-Stage Least Squares: 两阶段最小二乘回归。
- ⑪ Optimal Scaling: 最优尺度回归。

本章将详细介绍线性回归（Linear 过程）、曲线拟合（Curve Estimation 过程）、二分类变量 Logistic 回归（Binary Logistic 过程）和非线性回归（Nonlinear 过程）这 4 类最常用的回归分析方法。

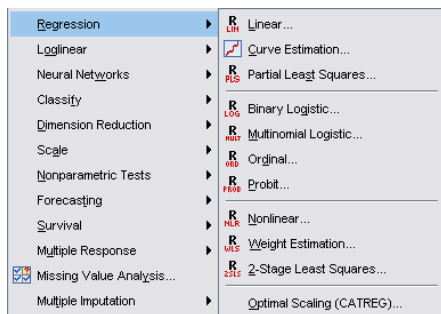


图 10-1 【Regression】子菜单

10.1.4 回归分析的一般步骤

在具体讲解各类回归分析方法之前，首先简单介绍一下回归分析的一般步骤。一个完整的回归分析通常包括以下几步。

STEP 01 对数据进行预处理，选择合适的变量进行回归分析。比如，要研究某地区一套家庭住宅的价格，则与之有关的变量就可能是房屋的面积、房龄、卧室数、邻居类型、住宅风格，等等。在建立回归模型的时候，选择哪些变量进入模型是首先要考虑的问题。对于变量的选取，既要考虑实际问题的背景也要考虑变量数据的统计特征。但是总的来说，实际问题远比数据的统计特性来得重要。

STEP 02 做散点图，观察变量间的趋势，初步选取回归分析方法。同时利用散点图剔除异常点。如图 10-2 所示，第一幅图中具有明显线性的趋势，所以应该选取线性回归的方法，但是其中有一个异常点，应该检查该点的数据是否存在观察或者录入错误应予以剔除。第二幅图则显然应该用非线性回归拟合变量之间的关系。

STEP 03 进行回归分析，拟合自变量与因变量之间的经验公式。

STEP 04 拟合完毕之后进行残差分析，检验模型是否恰当。设拟合经验公式计算的因变量估计值为 \hat{y} ，因变量实际取值为 y ，则 $y - \hat{y}$ 为残差值。残差分析主要包括检验残差是否独立，以及残差是否服从正态分布两方面。残差是否独立，通常采用 Durbin-Watson 残

差序列相关性检验来进行分析。残差分布是否为正态，可以用残差列表及一些相关指标来分析，但是最直观的方法为作残差的直方图。

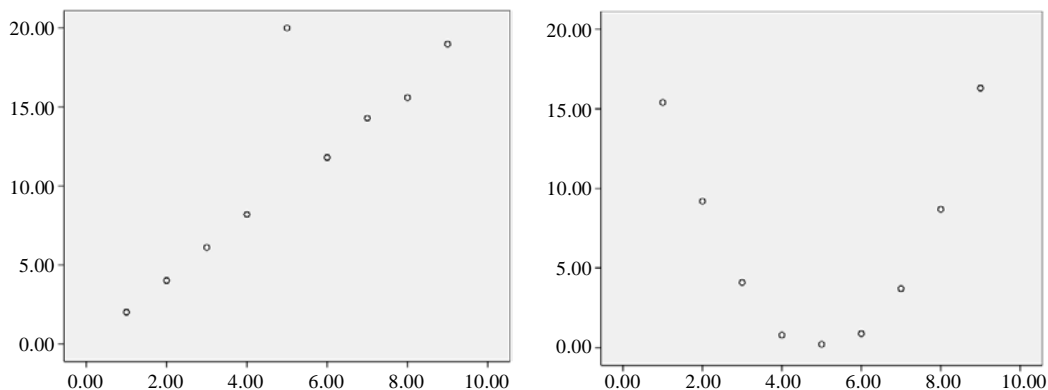


图 10-2 用散点图初步判定回归类型

STEP 05 利用拟合结果进行预测控制。

以上是进行回归分析的基本步骤。但是在处理实际问题的时候，一定要以问题的专业背景为基础，而不是拘泥于固定的数学方法。这也是统计学与传统数学的本质区别之一。

10.2 线性回归——Linear过程

线性回归在实际工作中发挥着重要作用。本节就将通过例子介绍线性回归的一般概念、操作界面和结果解释。

10.2.1 线性回归简介

根据自变量的个数，将线性回归分为一元线性回归和多元线性回归，两者既有很多共同点也有不同之处，下面将分别介绍。

1. 一元线性回归

一元线性回归模型是回归分析中处理两个变量间线性相关关系的最简单的数学模型。设变量 y 与 x 有下述关系：

$$y = a + bx + \varepsilon \quad (10.1)$$

其中， ε 是零均值的随机变量， x 是自变量， a, b 是未知参数，则称 (10.1) 为一元线性回归模型。通常在实际问题中，有 N 组训练数据，记为 $(x_i, y_i), (i = 1, 2, \dots, N)$ ，则有

$$y_i = a + bx_i + \varepsilon_i \quad (10.2)$$

其中 $\varepsilon_1, \varepsilon_2, \dots, \varepsilon_N$ 独立分布于 $N(0, \sigma^2)$ 。在回归模型 (10.1) 中，若能用某种方法求得未知参数 a, b 的估计 \hat{a}, \hat{b} ，则就可以估计因变量 y 的值如下：

$$\hat{y} = \hat{a} + \hat{b}x \quad (10.3)$$

称 (10.3) 式为一元线性回归方程或回归直线。在已知 x 的情况下，由 (10.3) 计算所得

的 \hat{y} 称为回归值。

利用最小二乘法, 则可由 (10.2) 中的训练数据得出 (10.3) 中 \hat{a} , \hat{b} 的估计值:

$$\hat{b} = \frac{l_{xy}}{l_{xx}}, \quad \hat{a} = \bar{y} - \hat{b}\bar{x} \quad (10.4)$$

其中:

$$l_{xy} = \sum_{i=1}^N (x_i - \bar{x})(y_i - \bar{y}) = \sum_{i=1}^N x_i y_i - N\bar{x}\bar{y}$$

$$l_{xx} = \sum_{i=1}^N (x_i - \bar{x})^2 = \sum_{i=1}^N x_i^2 - N\bar{x}^2$$

一般地, 对于任意一组观察值 $(x_i, y_i), (i=1, 2, \dots, N)$, 当 $l_{xx} \neq 0$ 时, 由 (10.4) 总可以求出回归直线 (10.3)。但是, 这样建立的回归方程是否有意义, 即 x 对 y 是否有影响, 且二者是否是线性关系, 还必须进行假设检验。

2. 多元线性回归

与一元线性回归模型不同, 多元线性回归模型是回归分析中处理一个因变量与多个自变量线性相关关系的数学模型。设变量 y 与 x_1, x_2, \dots, x_n 有下述关系:

$$y = b_0 + b_1 x_1 + b_2 x_2 + \dots + b_n x_n + \varepsilon \quad (10.5)$$

其中, ε 是零均值的随机变量, 则称 (10.5) 为多元线性回归模型。通常在实际问题中, 有 N 组训练数据, 记为 $(x_{i1}, x_{i2}, \dots, x_{in}, y_i), (i=1, 2, \dots, N)$, 则有

$$y_i = b_0 + b_1 x_{i1} + b_2 x_{i2} + \dots + b_n x_{in} + \varepsilon_i \quad (10.6)$$

其中 $\varepsilon_1, \varepsilon_2, \dots, \varepsilon_N$ 独立分布于 $N(0, \sigma^2)$ 。

还是利用最小二乘法, 我们可以求出对 $b_0, b_1, b_2, \dots, b_n$ 的估计值 $\hat{b}_0, \hat{b}_1, \hat{b}_2, \dots, \hat{b}_n$ 。与一元线性回归一样, 称

$$\hat{y} = \hat{b}_0 + \hat{b}_1 x_1 + \hat{b}_2 x_2 + \dots + \hat{b}_n x_n \quad (10.7)$$

为 n 元线性回归模型。这个模型是否恰当, 也需要进行假设检验。

3. 线性回归的适用条件

根据问题的不同, 线性回归的适用条件也不尽相同。高斯曾提出了 5 个假设理论, 满足这些假设条件的线性回归模型称为古典线性回归模型。这里也将其作为线性回归的基本适用条件。这些条件在介绍回归模型的时候也有所提及, 现系统地归纳如下:

- ① 正态性: 随机误差 $\varepsilon_1, \varepsilon_2, \dots, \varepsilon_N$ 均服从正态分布 $N(0, \sigma^2)$ 。
- ② 等方差性: 对于所有的 x_i, ε_i 的条件方差均为 σ^2 , 且 σ 为常数, 即 $\text{Var}(\varepsilon_i / x_i) = \sigma^2$ 。
- ③ 独立性: 即在给定 x_i 的条件下, ε_i 的条件期望均值为零。
- ④ 无自相关: 随机误差项 ε_i 的逐次观察值互不相关, 即 $\text{Cov}(\varepsilon_i, \varepsilon_j) = 0, (i \neq j)$ 。
- ⑤ ε 与 x 不相关: 随机误差项 ε_i 与相应的自变量 y_i 对因变量 y 的影响相互独立。换言之, 两者对因变量的影响是可以区分的, 即 $\text{Cov}(\varepsilon_i, x_i) = 0$ 。

以上 5 点可以作为线性回归模型的基本适用条件。在实际问题中, 如果只是建立方程, 而无需根据自变量的取值预测因变量的容许区间、可信区间等, 则前两个条件可以适当放宽。如果是小样本问题, 一般情况下也可以直接由散点图观察分析, 而无需特地对以上条

件进行一一判别。

10.2.2 Linear过程的操作界面

下面首先介绍【Linear】过程的操作界面。执行【Analyze】/【Regression】/【Linear】命令，弹出如图 10-3 所示的对话框。

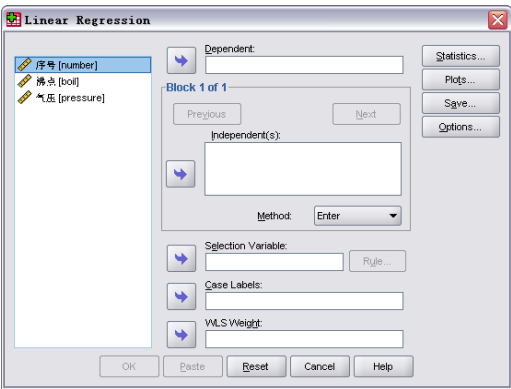


图 10-3 【Linear】对话框

该对话框主要由以下几部分组成。

1. 候选变量列表框

即左侧变量列表框。

2. 【Dependent】框

选择回归分析的因变量，只能选择一个。

3. 【Independent】框

选择回归分析的自变量，可以有一个或多个。当只有一个自变量时即进行一元线性回归，选择多个自变量时进行多元线性回归。

4. 【Method】下拉列表

多元线性回归中定义【Independent】框中自变量进入模型的方式。各个方式的具体说明如表 10-2 所示。

表 10-2 【Method】下拉列表对应的具体方法

SPSS 中表示	中 文 名	说 明
Enter	强行进入法	将【Independent】框中自变量全部纳入回归模型中，不做任何筛选
Stepwise	逐步法	根据【Options】对话框中设定的条件逐个选取变量进入模型之中。具体的选取办法是首先计算各个自变量对因变量的影响大小。选取影响最大的变量进入模型之中 然后重复此过程，需要注意此时新变量的引入是否会使先前变量丧失统计意义。如果会的话，这个变量就要予以剔除并重新计算剩余变量对因变量的影响大小。直至方程中没有可剔除的变量，方程外没有变量可以引入为止

续表

SPSS 中表示	中 文 名	说 明
Remove	强制剔除法	只出不进，根据移出标准将同一个 Block 里边的变量一次全部剔除
Backward	向后剔除法	筛选方法和逐步法类似。但是只出不进，即对选入变量按照对因变量的影响大小由小到大依次剔除，直到所有变量均符合选入标准为止
Forward	向前剔除法	与向后剔除法相反，只进不出。对于已纳入方程的变量就不再考虑其显著性，直到方程外变量都达不到入选标准，没有新变量可以引入为止

注意 变量的选择不是单纯的数学问题，一定要结合实际问题的背景来处理。这一点也是本书中反复强调的地方。也就是说一定要认识到统计学和传统数学的不同。传统数学是一门要求严密逻辑推理的学科。统计学作为一门从应用中发展起来的学科，一定不能脱离实际问题。否则，统计学也就丧失了其存在的意义。

5. 【Selection Variable】框

选择筛选变量。需要注意的是筛选变量不能重复选择其他变量选择框已经选入的变量。当【Selection Variable】框选入一个变量之后，则激活其后的【Rules】按钮。单击【Rules】按钮，弹出如图 10-4 所示的【Rules】对话框。该对话框主要用于给定变量的筛选条件。只有满足该条件的记录才进入回归分析。



图 10-4 【Rules】对话框

6. 【Case Labels】框

选择变量作为每条记录的标签，通常选取记录号。

7. 【WLS Weight】框

选择权重变量进行加权最小二乘回归分析。在分析时按照权重变量的大小给每条记录赋予不同的权重值。

8. 【Statistics】

单击图 10-3 中的【Statistics】按钮，弹出如图 10-5 所示的【Statistics】对话框。

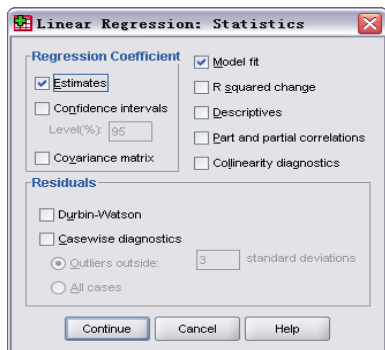


图 10-5 【Statistics】对话框

该对话框用于选择需要计算的统计量，主要由以下几部分组成。

- **Regression Coefficients** 复选框组

定义回归系数的输出情况，各项的具体作用如表 10-3 所示。

表 10-3 Regression Coefficients 复选框组

英文名	作用	备注
Estimation	输出回归系数的估计值及其标准误差、检验统计量，标准化的回归系数	系统默认选项
Confidence intervals	输出每个回归系数 95% 置信区间	——
Covariance Matrix	输出每个自变量的相关矩阵、方差和协方差矩阵	——

- **Model fit** 复选框

选中后输出回归模型因变量列表和模型是否恰当的一些检验统计量，以及复相关系数 R 、决定系数 R^2 和调整的 R^2 、方差分析表等。此项为系统默认选项。

- **R squared change** 复选框

选中后输出模型拟合过程中 R^2 、F 值和 P 值的改变情况。

- **Descriptives** 复选框

选中后输出描述性统计量。

- **Part and partial correlations** 复选框

选中后输出自变量间的相关系数、部分相关系数和偏相关系数。

- **Collinearity diagnostics** 复选框

选中后输出多元线性回归中用于共线性诊断的统计量。

- **Residuals** 复选框

输出残差分析的结果。

① **Durbin-Watson** 复选框：选中后输出 Durbin-Watson 残差序列相关性检验结果。

② **Casewise diagnostics**：选中后输出超过规定 n 倍标准差的残差列表或全部残差列表。

9. 【Plots】

单击图 10-3 中的【Plots】按钮，弹出如图 10-6 所示的【Plots】对话框。

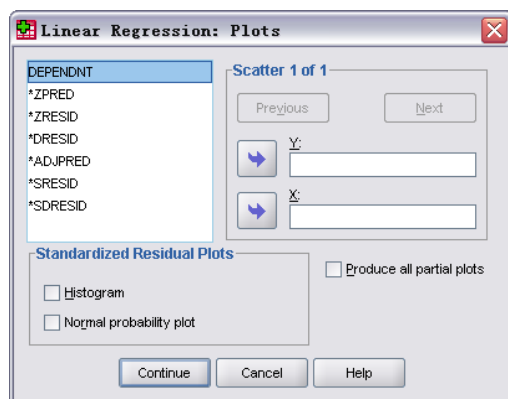


图 10-6 【Plots】对话框

该对话框主要用于绘制各类图形，具体包括以下几项。

① 候选变量框：列举出可以用来绘制图形的中间统计量，包括因变量（DEPENDNT）、标准化预测值（ZPRED）、标准化残差（ZRESID）、剔除残差（DRESID）、修正后预测值（ADJPRED）、学生化残差（SRESID）和学生化剔除残差（SDRESID）。

② Scatter 组：从左侧候选变量框中选择变量到 X、Y 轴框，定义需要绘制的回归分析诊断图或预测图。

③ Standardized Residual Plots 复选框组：选择绘制标准化残差图的类型，包括直方图（Histogram）和正态 P-P 图（Normal probability plot）。

④ Produces all partial plots 复选框：选择是否绘制每一个自变量与因变量残差的散点图。

10. 【Save】

单击图 10-3 中的【Save】按钮，弹出如图 10-7 所示的【Save】对话框。

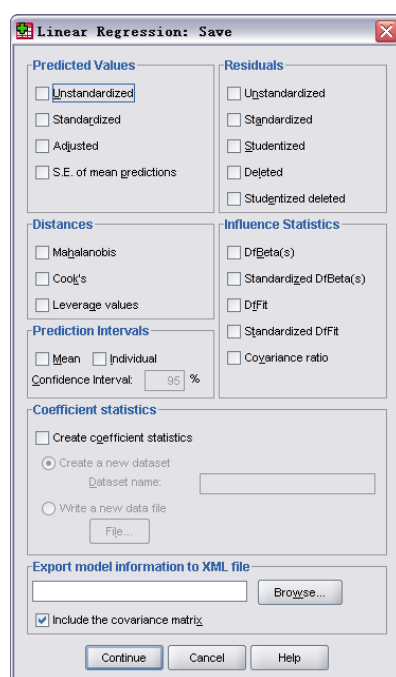


图 10-7 【Save】对话框

【Save】对话框主要用来存储各个分析的中间结果。该对话框主要包括以下选项。

① Predicted Values 复选框组：如表 10-4 所示，主要用于保存预测值。

表 10-4 Predicted Values 复选框组各项具体作用

SPSS 中表示	作 用
Unstandardized	保存模型对因变量的原始预测值
Standardized	保存标准化后的预测值，此时均值为 0，标准差为 1
Adjusted	保存去掉当前记录时，当前模型对该记录因变量的预测值
S.E of mean predictions	保存预测值的标准差

② Residuals 复选框组：如表 10-5 所示，保存用于回归诊断时所需的各种残差。

表 10-5 Residuals 复选框组各项具体作用

SPSS 中表示	作 用
Unstandardized	保存模型预测值对因变量观测值的原始残差
Standardized	保存用 U 变换进行标准化后的残差，此时均值为 0，标准差为 1
Studentized	保存学生化残差，即用 T 变换进行标准化后的残差
Delete	保存删除当前记录后的残差
Studentized Delete	保存删除当前记录后，用 T 变换进行标准化后的残差

③ Distances 复选框组：如表 10-6 所示，保存用于测量数据点离拟合模型距离的指标，通常用于诊断离群点或强影响点。

表 10-6 Distances 复选框组各项具体作用

SPSS 中表示	作 用
Mahalanobis	马氏距离，保存记录值离样本平均值的距离
Cook's	保存删除当前记录后，模型残差会发生的变化量
Leverage values	杠杆值，测量该数据点的影响强度

④ Influence statistics 复选框组：如表 10-7 所示，保存用于判断强影响点的统计量。

表 10-7 Influence statistics 复选框组各项具体作用

SPSS 中表示	作 用	备 注
DfBeta	保存去掉该观察值后回归系数的变化值	
Standardized DfBeta	保存标准化的 DfBeta 值	当其大于 $2/\sqrt{N}$ 时，该点可能为强影响点，其中 N 表示样本个数
DfFit	保存去掉该观测点后预测值的变化值	
Standardized DfFit	保存标准化的 DfFit	当其大于 $2/\sqrt{P/N}$ 时，该点可能为强影响点。其中 P 为变量数
Covariance ratio	保存去掉该观测点之后协方差阵与含全部观测值的协方差阵的比率	当其绝对值大于 $3P/N$ 时，该观测值可能为强影响点

⑤ Prediction intervals 复选框组：选择是否给出均值和个体参考值的置信区间。

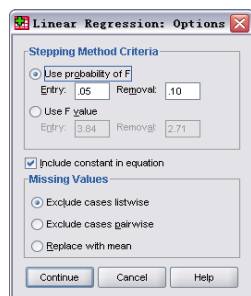
⑥ Coefficient statistics 组：该组主要用于保存上述中间变量。SPSS 17.0 提供了两种保存方法，可以将结果保存在一个新生成的“*.sav”数据文件中（create a new data set），也可以将结果直接保存到其他文件里（write a new data file）。

⑦ Export model information to XML file：将模型信息存入 XML 文件中。

11. 【Options】

单击图 10-3 中的【Options】按钮，弹出如图 10-8 所示【Options】

图 10-8 【Options】对话框 对话框。



该对话框主要用来设置回归分析的一些选项。包括以下几项。

- **Stepping Method Criteria** 组

设置变量纳入和排除标准。

- **Include constant in equation** 复选框

用于决定模型中是否包括常数项，默认选中。

- **Missing Values** 单选框组

定义缺失值的处理方式。

① **Exclude cases listwise**: 只要数据中有变量值缺失就剔除该数据。

② **Exclude cases pairwise**: 仅当数据要分析的变量值缺失时才剔除该数据。

③ **Replace with mean**: 用变量均值代替变量缺失值。

10.2.3 一元线性回归的例子

下面通过例子介绍【Linear】过程及其结果解释。

例 10.1 在十九世纪四五十年代，苏格兰物理学家 James D.Forbes，试图通过水的沸点来估计海拔高度。由于可以通过气压来估计海拔，他在阿尔卑斯山及苏格兰收集了沸点及海拔的数据如表 10-8 所示。现在通过线形回归拟合气压与沸点的关系。（数据来源：《应用线性回归》，中国统计出版社）。

表 10-8 Forbes 数据

序 号	沸 点	气压（英寸汞柱）	序 号	沸 点	气压（英寸汞柱）
1	194.5	20.79	10	201.3	24.01
2	194.3	20.79	11	203.6	25.14
3	197.9	22.40	12	204.6	26.57
4	198.4	22.67	13	209.5	28.49
5	199.4	23.15	14	208.6	27.76
6	199.9	23.35	15	210.7	29.04
7	200.9	23.89	16	211.9	29.88
8	201.1	23.99	17	212.2	30.06
9	201.4	24.02			

这是一个一元线性回归问题，执行以下操作。

STEP 01 建立如图 10-9 所示的数据文件“Forbes.sav”。其中“boil”表示沸点，“pressure”表示气压。

	number	boil	pressure
1	1.00	194.50	20.79
2	2.00	194.30	20.79
3	3.00	197.90	22.40

图 10-9 数据文件“Forbes.sav”的数据结构

STEP 02 首先通过做沸点与气压之间的散点图判断两者的关系。如图 10-10 所示，两者有明显的线性关系。

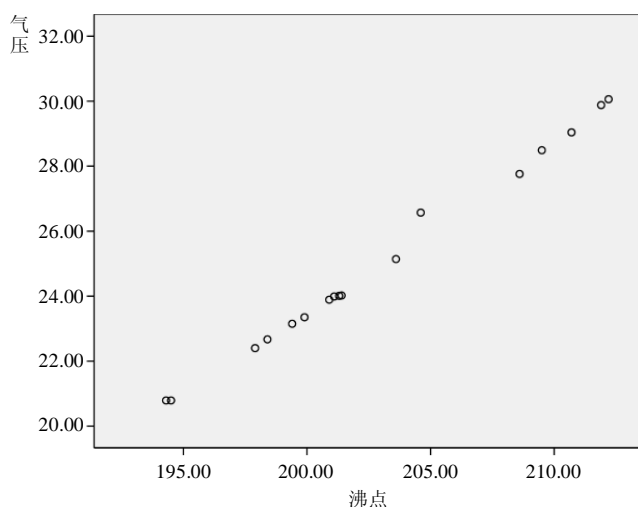


图 10-10 沸点与气压的散点图

STEP 03 一元线性回归分析，执行以下操作：

执行【Analyze】/【Regression】/【Linear】命令，弹出【Linear】对话框

【Dependent】框：pressure

定义气压为因变量

【Independent】框：boil

定义沸点为自变量

单击【Plots】按钮

弹出【Plots】对话框

【Plots】对话框：

Standardized Residual Plots 复选框组：

选中“Histogram”和“Normal probability plot”复选框，输出残差直方图和正态 P-P 图

单击【Continue】按钮

【Plots】对话框定义完成

单击【OK】按钮

定义完成

执行以上操作之后，生成表 10-9～表 10-13。

表 10-9 表示回归分析过程中变量进入、退出模型的基本情况。这张表主要是针对多元线性回归的情况。此处可以忽略。

表 10-9 变量记录情况

Variables Entered/Removed^b

Model	Variables Entered	Variables Removed	Method
1	沸点 ^a	.	Enter

a. All requested variables entered.

b. Dependent Variable: 气压

表 10-10 表示回归模型的拟合度。其中第二列 R 表示复相关系数，其含义类似于上一章所讲的相关系数，反映的是自变量与因变量之间的密切程度。其值在 0 到 1 之间，越大越好。第三列 R^2 是复相关系数的平方，又称为决定系数。但是需要注意的是，复相关系

数随着模型中自变量个数的增加，其值是不断增大的。所以，对于多元线性回归模型复相关系数就不太可靠。于是又引入了调整的复相关系数。表 10-10 的第四列即表示调整后的决定系数，是在考虑了模型中自变量个数的情况下计算的决定系数。在一元线性回归的时候，其值等于决定系数的值。第五列表示标准误。通过观察这几个数据可知，本例的拟合情况是很好的。

表 10-10 模型拟合度检验

Model Summary^b

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.997 ^a	.994	.994	.23283

a. Predictors: (Constant), 沸点

b. Dependent Variable: 气压

表 10-11 是模型检验结果。这是一个标准的方差分析表。回归模型的 Sig. 值为 0，说明该模型有显著的统计意义。

表 10-11 方差分析表

ANOVA^b

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	145.125	1	145.125	2677.105	.000 ^a
	Residual	.813	15	.054		
	Total	145.938	16			

a. Predictors: (constant), 沸点

b. Dependent Variable: 气压

表 10-12 给出了拟合未标准化的和标准化之后的回归系数值（含常数项），并通过 t 检验方法对拟合结果进行检验，常数项和沸点所对应的系数其 t 检验的 Sig. 值都为 0，具有显著的统计学意义。

表 10-12 回归分析结果

Coefficients^a

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	-81.064	2.052		-39.508	.000
	沸点	.523	.010	.997	51.741	.000

a. Dependent Variable: 气压

根据表 10-12，得出例 10-1 的拟合结果为 $pressure = 0.523boil - 81.064$ 。

虽然通过表 10-12 拟合得出了沸点与气压间的经验公式，但是一个完整的回归分析过程还包括利用残差分析，对拟合结果进行检验。表 10-13 所示是与残差有关的一些统计量，包括预测值及标准化的预测值、残差及残差的预测值的最小值、最大值、均值、标准差和

样本数。这些数据中无离群值，且数据的标准差也比较小，可以认为模型是健康的。

表 10-13 残差统计量

	Minimum	Maximum	Mean	Std. Deviation	N
Predicted Value	20.5343	29.8940	25.0588	3.01169	17
Residual	-.25717	.64994	.00000	.22544	17
Std. Predicted Value	-1.502	1.605	.000	1.000	17
Std. Residual	-1.105	2.791	.000	.968	17

a. Dependent Variable: 气压

对于模型的检验，除了分析残差统计量外，还可以直接做出标准化残差的直方图和正态 P-P 图来观察其是否服从正态分布。

通过观察如图 10-11 所示的标准化残差直方图和如图 10-12 所示的标准残差正态 P-P 图可以发现，由于残差具有正态分布的趋势。因此可以认为这里的回归模型是恰当的。

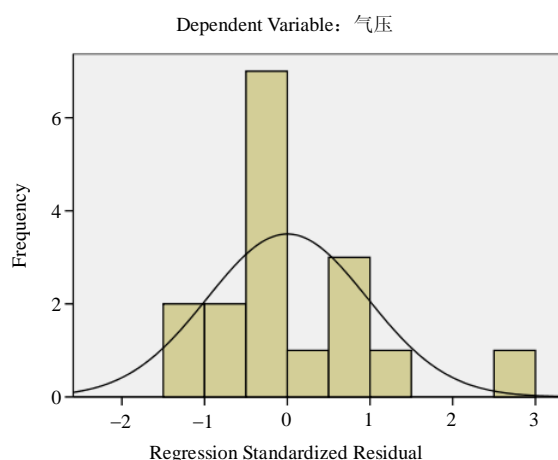


图 10-11 标准化残差直方图

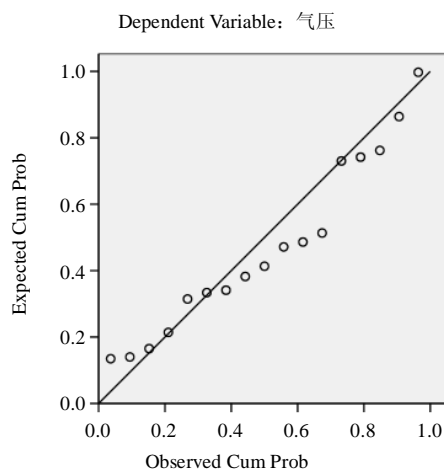


图 10-12 标准化残差正态 P-P 图

10.2.4 多元线性回归的例子

下面再给出一个【Linear】过程在多元线性回归中的例子。

例 10.2 某大型金融机构中做了一项关于雇员对其主管满意度的调查，其中一个问题设计为对主管的工作业绩的综合评价，另外若干个问题涉及主管与其雇员间相互关系的具体方面。该研究试图解释主管性格与雇员对其整体满意度之间的关系。起初选取了 6 个调查项目作为可能的解释变量，表 10-14 给出了这些变量。

表 10-14 主管人员业绩数据的变量描述

变 量	定 义
Y	对主管工作情况的总体评价
X1	处理雇员的抱怨

续表

变 量	定 义
X2	不允许特权
X3	学习新知识的机会
X4	根据工作业绩升职
X5	对不良表现过于吹毛求疵
X6	提升到更好工作的速度

试通过多元线性回归的方法分析表 10-15 中的数据来研究这个问题。(数据来源:《例解回归分析》, 中国统计出版社)

表 10-15 主管人员业绩数据

序 号	Y	X1	X2	X3	X4	X5	X6
1	43.00	51.00	30.00	39.00	61.00	92.00	45.00
2	63.00	64.00	51.00	54.00	63.00	73.00	47.00
3	71.00	70.00	68.00	69.00	76.00	86.00	48.00
4	61.00	63.00	45.00	47.00	54.00	84.00	35.00
5	81.00	78.00	56.00	66.00	71.00	83.00	47.00
6	43.00	55.00	49.00	44.00	54.00	49.00	34.00
7	58.00	67.00	42.00	56.00	66.00	68.00	35.00
8	71.00	75.00	50.00	55.00	70.00	66.00	41.00
9	72.00	82.00	72.00	67.00	71.00	83.00	31.00
10	67.00	61.00	45.00	47.00	62.00	80.00	41.00
11	64.00	53.00	53.00	58.00	58.00	67.00	34.00
12	67.00	60.00	47.00	39.00	59.00	74.00	41.00
13	69.00	62.00	57.00	42.00	55.00	63.00	25.00
14	68.00	83.00	83.00	45.00	59.00	77.00	35.00
15	77.00	77.00	54.00	72.00	79.00	77.00	46.00
16	81.00	90.00	50.00	72.00	60.00	54.00	36.00
17	74.00	85.00	64.00	69.00	79.00	79.00	63.00
18	65.00	60.00	65.00	75.00	55.00	80.00	60.00
19	65.00	70.00	46.00	57.00	75.00	85.00	46.00
20	50.00	58.00	68.00	54.00	64.00	78.00	52.00
21	50.00	40.00	33.00	34.00	43.00	64.00	33.00
22	64.00	61.00	52.00	62.00	66.00	80.00	41.00
23	53.00	66.00	52.00	50.00	63.00	80.00	37.00
24	40.00	37.00	42.00	58.00	50.00	57.00	49.00
25	63.00	54.00	42.00	48.00	66.00	75.00	33.00
26	66.00	77.00	66.00	63.00	88.00	76.00	72.00
27	78.00	75.00	58.00	74.00	80.00	78.00	49.00
28	48.00	57.00	44.00	45.00	51.00	83.00	38.00
29	85.00	85.00	71.00	71.00	77.00	74.00	55.00
30	82.00	82.00	39.00	59.00	64.00	78.00	39.00

STEP 01 选择模型中的变量。

建立数据文件“supervisor.sav”。首先分析各个待选变量的特征。在这 6 个解释变量中有两个主要类型：变量 X1、X2 和 X5 反映的是雇员与主管人员之间直接的人际关系，X3 和 X4 主要和工作有关。变量 X6 不是对主管的直接评价，而是雇员对自己把握晋升机会的一般评价。因此，在进行多元线性回归的时候将 X1、X2、X5 直接纳入模型，X3 和 X4 通过逐步法进入模型。而 X6 直接不予考虑。

STEP 02 多元回归分析。执行以下操作：

执行【Analyze】/【Regression】/【Linear】命令，弹出【Linear】对话框	
【Dependent】框：Y	定义对主管工作情况的总体评价为因变量
【Independent】框：X1、X2、X5	定义第一组自变量，进入方式为“Enter”
单击【Next】按钮	定义下一组自变量
【Independent】框：X3、X4	定义第二组自变量
【Method】下拉列表：stepwise	定义第二组自变量进入方式为“Stepwise”
单击【Statistics】按钮	弹出【Statistics】对话框
【Statistics】对话框：	
选中“Estimates”和“Model fit”复选框	系统默认选项
选中 Collinearity diagnostics 复选框	选择对多元回归进行共线性诊断
单击【Continue】按钮	【Statistics】对话框定义完成
单击【OK】按钮	定义完成

执行以上操作之后生成如表 10-16～表 10-20 所示的 5 张表格，分别解释如下。

表 10-16 是变量进入模型的基本情况。在本例的操作中，将所有变量分为两组。其中 X1、X2、X5 分为一组，采用强行进入法纳入模型。X3、X4 分为一组，采用逐步法进入模型。但从表 10-16 中可以看出，最后只建立一个模型，即将 X1、X2、X5 全部纳入模型，而 X3、X4 全部剔除。

表 10-16 变量进入情况

Variables Entered/Removed^b

Model	Variables Entered	Variables Removed	Method
1	对不良工作表现过于吹毛求疵 X5, 不允许特权 X2, 处理雇员的抱怨 X1 ^a	.	Enter

a. All requested variables entered.

b. Dependent Variable: 对主管工作情况的总体评价 Y

表 10-17 所示是对模型拟合度的检验结果。对于多元线性回归模型，一般应采用其调整的决定系数（Adjusted R Square）来判断。在本例中，其值为 0.647，说明其拟合程度还是可以接受的。

表 10-17 模型拟合度检验

Model Summary				
Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.826 ^a	.683	.647	7.23720

a. Predictors: (Constant), 对不良工作表现过于吹毛求疵 X5, 不允许特权 X2, 处理雇员的抱怨 X1

表 10-18 所示是模型检验结果。这是一个标准的方差分析表。回归模型的 Sig. 值为 0, 说明该模型有显著的统计意义。

表 10-18 方差分析表

ANOVA ^b						
Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	2935.162	3	978.387	18.680	.000 ^a
	Residual	1361.805	26	52.377		
	Total	4296.967	29			

a. Predictors: (Constant), 对不良工作表现过于吹毛求疵 X5, 不允许特权 X2, 处理雇员的抱怨 X1

b. Dependent Variable: 对主管工作情况的总体评价 Y

表 10-19 所示是回归分析的结果。表格从左到右依次表示未标准化的回归系数 (Unstandardized Coefficients)、标准化的回归系数 (Standardized Coefficients)、t 检验统计量值、Sig. 值和共线性检验统计量值 (Collinearity Statistics)。

表 10-19 回归分析结果

Coefficients ^a								
Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Collinearity Statistics	
		B	Std. Error	Beta			Tolerance	VIF
1	(Constant)	15.028	11.536		1.303	.204		
	处理雇员的抱怨 X1	.780	.123	.853	6.357	.000	.677	1.477
	不允许特权 X2	-.050	.133	-.051	-.380	.707	.686	1.457
	对不良工作表现、对 于吹毛求疵 X5	.005	.138	.004	.034	.973	.962	1.039

a. Dependent Variable: 对主管工作情况的总体评价 Y

由未标准化的回归系数可知, 本例的拟合结果为 $Y=0.78X_1-0.05X_2+0.005X_3+15.028$ 。从 Sig. 取值可知, 仅 X_1 的系数是有统计学意义的。对于多元线性模型, 通常还应检验其自变量之间是否存在共线性的问题。通常检验共线性有如表 10-20 所示的几个指标。

表 10-20 共线性检验指标

指标名称	检验标准
容忍度 (Tolerance)	若某自变量容忍度小于 0.1, 则存在共线性问题

续表

指标名称	检验标准
方差膨胀率（VIF）	容忍度的倒数越大共线性问题越严重
特种根（Eigenvalue）	若多个维度的特征根等于 0，则可能存在共线性问题
条件指数（Condition Index）	若某个维度的条件指数大于 30，则可能存在共线性问题

本例的模型中不存在共线性问题。如果模型中存在共线性问题，那么就应当通过增大样本量或重新建立模型来解决此问题。

表 10-21 所示是未进入模型的变量列表。这两个变量的 Sig.取值均大于 0.05，说明在模型中无需再对其进行分析了。

表 10-21 剔除变量列表

Excluded Variables^b

Model	Beta In	t	Sig.	Partial Correlation	Collinearity Statistics		
					Tolerance	VIF	Minimum Tolerance
1 学习新知识的机会 X3	.230 ^a	1.674	.107	.318	.607	1.649	.544
依据工作业绩升职 X4	.084 ^a	.521	.607	.104	.481	2.078	.464

a. Predictors in the Model: (Constant), 对不良工作表现过于吹毛求疵 X5, 不允许特权 X2, 处理雇员的抱怨 X1

b. Dependent Variable: 对主管工作情况的总体评价 Y

表 10-22 所示是模型的共线性检验结果。由表中数据并结合表 10-20 的检验标准可知，本例中模型不存在共线性的问题。

表 10-22 共线性检验

Collinearity Diagnostics^a

Model	Dimension	Eigenvalue	Condition Index	Variance Proportions			
				(Constant)	处理雇员的抱怨 X1	不允许特权 X2	对不良工作表现、对于吹毛求疵 X5
1	1	3.939	1.000	.00	.00	.00	.00
	2	.035	10.627	.06	.06	.43	.14
	3	.018	14.782	.01	.92	.56	.02
	4	.008	22.086	.93	.02	.01	.84

a. Dependent Variable: 对主管工作情况的总体评 Y

10.2.5 小结

本节结合例子介绍了一元线性回归和多元线性回归的一般概念及其在 SPSS 中的实现。对于多元线性回归，其变量的选取和最后模型的检验都是十分复杂的问题。这些问题既依赖于统计学知识，同时又更应该与实际问题的具体背景结合起来。当两者发生差异时，则依赖于实际工作者的经验。希望读者能慢慢体会到这一点。

10.3 曲线拟合——Curve Estimation过程

上一节介绍了线性回归的概念及其在 SPSS 中的实现。本节将通过例子介绍两变量间的曲线拟合。

10.3.1 曲线拟合简介

在实际问题中，变量间的关系可以是线性的，也可以是非线性的。若变量间存在着线性关系，那么可以用线性回归的方法来拟合两者之间的关系。若变量间的关系是非线性的，那么问题就变得复杂得多。在线性与非线性之间，还有一种拟线性关系。所谓拟线性，是指变量之间的关系是非线性关系，但是可以通过一些特殊的变化使之线性化。比如函数 $y = ce^x$ ， y 与 x 是非线性关系。但是对方程两边同时取自然对数，原问题变为 $\ln y = x + \ln c$ ，此时 $\ln y$ 与 x 即为线性关系。

曲线拟合就是研究两变量间拟线性关系的一种方法。曲线拟合的基本步骤是首先选择一种常见的曲线模型及其数学表达式，然后对变量做变换使得曲线模型线性化，再利用已知数据，用最小二乘的方法来估计模型中的参数。

利用曲线拟合的方法估计两变量间的关系，必须选取恰当的曲线模型。模型的选取首先依赖于实际问题，比如要估计某封闭地区人口增长与时间的关系，那么可以用 Growth 曲线来拟合二者的关系，但同时也依赖于数据的特征。这一点可以通过做变量间的散点图来对两变量间的关系进行一个预估计。再根据预估计结果选择恰当的统计模型。

需要注意的是，由于 SPSS 的【Curve Estimation】过程所提供的曲线模型是十分有限的。因此，通常还需要对曲线拟合的结果做进一步的分析。

10.3.2 Curve Estimation过程的操作界面

下面首先介绍【Curve Estimation】过程的操作界面。执行【Analyze】/【Regression】/【Curve Estimation】命令，弹出如图 10-13 所示对话框。



图 10-13 【Curve Estimation】对话框

该对话框主要由以下几部分组成。

1. 候选变量列表框

即左侧变量列表框，给出了所有可以用曲线拟合的变量。

2. 【Dependent】框

定义曲线拟合的因变量即图形的 Y 轴。若选择多个变量，则分别拟合各个因变量与自变量间的关系。

3. 【Independent】组

定义曲线拟合的自变量即图形的 X 轴。只能选择一个变量，自变量主要有以下两种形式。

- ① 【Variable】框：选择候选变量列表框中的一个变量作为自变量。
- ② “Time” 单选框：选择时间作为自变量，此时所选择的因变量必须是时间序列数据。

4. 【Case Labels】框

定义图形中的标识变量。

5. “include constant in equation” 复选框

选择曲线模型的方程中是否包含常数项，系统默认选中此项。

6. “Plot models” 复选框

选择是否绘制拟合曲线的图形，系统默认选中此项。

7. 【Models】复选框组

定义拟合的曲线模型。SPSS 一共提供了如表 10-23 所示的 11 种模型。

表 10-23 曲线拟合的模型

SPSS 中表示	拟合曲线名	拟合曲线的数学表达式
Linear	直线	$y = b_0 + b_1x$
Quadratic	二次曲线	$y = b_0 + b_1x + b_2x^2$
Compound	复合曲线	$y = b_0 \cdot b_1^x$
Growth	生长曲线	$y = e^{b_0 + b_1x}$
Logarithmic	对数曲线	$y = b_0 + b_1 \ln x$
Cubic	三次曲线	$y = b_0 + b_1x + b_2x^2 + b_3x^3$
S	S 曲线	$y = e^{b_0 + b_1 / x}$
Exponential	指数曲线	$y = b_0 e^{b_1x}$
Inverse	逆变换曲线	$y = b_0 + b_1 / x$
Power	乘幂曲线	$y = b_0 \cdot x^{b_1}$
Logistic	Logistic 曲线	$y = \frac{1}{1/u + b_0 e^x}$ (u 为待定参数)

8. “Display ANOVA table” 复选框

选择是否输出曲线拟合模型检验的方差分析表。

9. 【Save】

单击【Save】按钮，弹出如图 10-14 所示对话框。

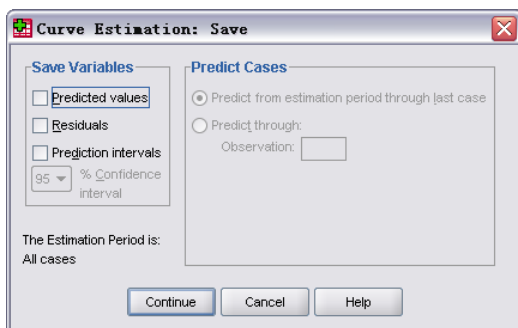


图 10-14 【Save】对话框

该对话框主要用于定义需要保存的统计量。包括以下两部分。

- Save Variables 复选框组

定义需要保存的中间统计量。

① Predicted Values 复选框：选择是否保存预测值。

② Residuals 复选框：选择是否保存残差。

③ Predicted intervals 复选框：选择是否保存预测值的置信区间。系统默认置信度为 95%。

- Predict Cases 单选框组

定义预测观测值组。仅当在【Independent】组中选择“Time”单选框组时该组才被激活。

① Predict from estimation period through last case 单选框：对估计周期内的所有观测值估计其预测值。这个周期可以由【Data】菜单中的【Select Cases】过程定义。如果不定义则输出全部观测值的预测值。

② Predict through 单选框：预测时间序列中最后一个观测值之后的 n 个值。 n 值可以由该单选框下方的 Observation 框定义。

- The Estimation Period is

显示当前的估计周期。

10.3.3 引例及结果解释

下面通过例子来介绍【Curve Estimation】过程的实现。

例 10.3 已知某次泥石流的各阵观测数据保存在如图 10-15 所示的数据文件“nishiliu.sav”中，试拟合各阵泥石流泥面宽与泥深之间的关系。（数据来源：《云南蒋家沟泥石流运动资料观测集》，科学出版社）

	序号	流态	龙头时间	龙尾时间	历时	泥面宽	泥深
1	8.00	阵性流S	3:11:55	3:12:08	13.00	20.00	0.80
2	9.00	阵性流S	3:12:58	3:13:05	7.00	20.00	0.80
3	10.00	阵性流S	3:15:21	3:15:55	3.00	35.00	2.00
4	11.00	阵性流S	3:16:57	3:17:32	55.00	30.00	1.50

图 10-15 数据文件“nishiliu.sav”的数据结构

通过【Curve Estimation】过程来拟合二者之间的关系，执行以下操作。

STEP 01 做散点图，分析二者之间的关系。散点图如图 10-16 所示。

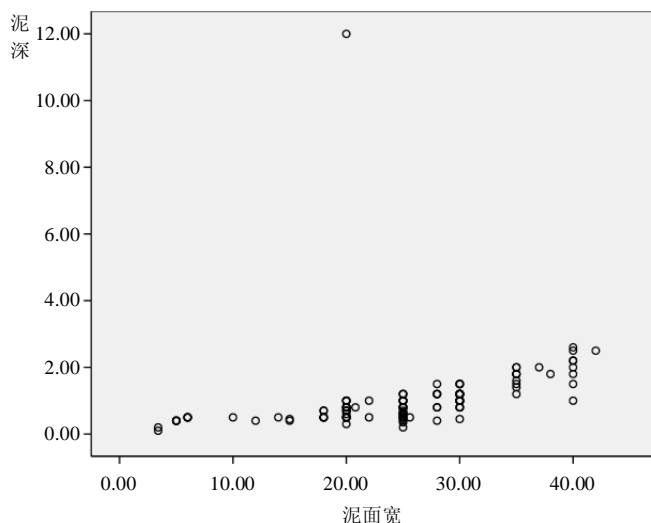


图 10-16 泥面宽与泥深的关系散点图

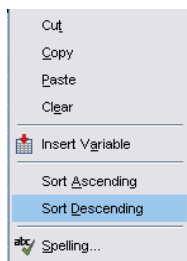


图 10-17 右键弹出菜单

通过观察散点图可以发现，图形中有一个明显的异常点。结合实际经验判断该值极有可能是录入错误，有必要剔除该数据。

STEP 02 剔除异常值。从图形上看，异常值应该是变量“泥深”的数据录入有误。因此在数据视图窗中选择变量名“泥深”，单击鼠标右键，弹出如图 10-17 所示的右键弹出菜单。在该菜单中选择【Sort Descending】过程，则所有记录将会按照泥深大小从大到小排序。此时异常值所在记录就会排到第一行，予以剔除。

STEP 03 曲线拟合。执行以下操作：

执行【Analyze】/【Regression】/【Curve Estimation】命令，弹出【Curve Estimation】对话框	
【Dependent】框：泥深	定义为泥深因变量
【Independent】组：	
选择【Variable】框：泥面宽	定义泥面宽为自变量
选中“include constant in equation”复选框	定义模型中包括常数项

选中“Plot models”复选框	定义绘制模型的曲线
选中“Linear”、“Quadratic”、“Cubic”复选框	分别用直线、二次曲线、三次曲线模型拟合
选中“Display ANOVA table”复选框	输出拟合模型的检验
单击【OK】按钮	定义完成

执行以上操作之后，会在 SPSS 的结果浏览窗口中生成多张图表。前面的图表是对模型基本信息的描述，这里就略过不讲了。在本例中，共选择了 3 个模型。每个模型都会生成如表 10-24～表 10-26 所示的 3 张表格。下面以二次曲线为例，分别介绍这 3 张表格。

表 10-24 是对二次曲线模型的检验结果。从表格上看，其调整的决定系数值为 0.733，说明模型的拟合效果不错。

表 10-24 二次曲线模型的模拟度检验

Model Summary			
R	R Square	Adjusted R Square	Std. Error of the Estimate
.859	.737	.733	.276

The independent variable is 泥面宽.

表 10-25 所示是二次曲线模型的方差分析表。其 Sig.取值为 0，说明模型具有显著的统计学意义。

表 10-25 二次曲线模型的方差分析表

ANOVA					
	Sum of Squares	df	Mean Square	F	Sig.
Regression	24.734	2	12.367	162.762	.000
Residual	8.814	116	.076		
Total	33.547	118			

The independent variable is 泥面宽.

表 10-26 是二次曲线的系数。从表格中可知 $y = 0.54 - 0.032x + 0.002x^2$ ，其中 y 表示泥深， x 表示泥面宽。从各系数的 Sig.值可以看出，各项系数都是有显著意义的。

表 10-26 二次曲线模型的系数

Coefficients					
	Unstandardized Coefficients		Standardized Coefficients	t	Sig.
	B	Std. Error	Beta		
泥面宽	-.032	.011	-.532	-2.937	.004
泥面宽**2	.002	.000	1.361	7.510	.000
(Constant)	.540	.124		4.342	.000

对比表 10-24～表 10-26 可以看出，曲线拟合的输出结果和线性回归的输出结果是十

分类似的。若选择 Linear 模型进行曲线拟合就等同于对两变量进行线性回归分析。

图 10-18 所示是拟合曲线和原始观测值的图形。从图形上看，3 个模型的拟合效果是比较近似的。

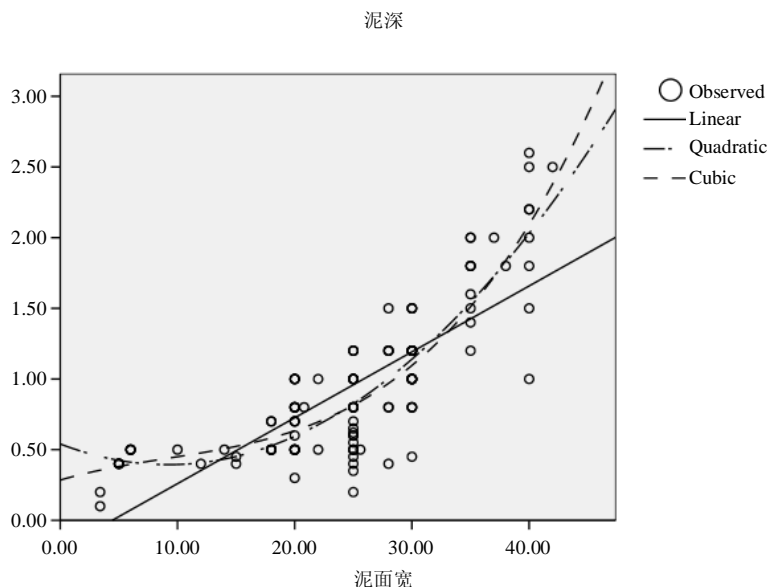


图 10-18 曲线拟合结果

10.4 二分类变量 Logistic 回归——Binary Logistic 过程

前面所介绍的线性回归和曲线拟合都要求因变量为定量变量。但是在实际问题中因变量却是既有定量变量，也有定性变量。对于因变量是定性变量的情况，引入了 Logistic 回归的方法。本节通过例子讲解二分类变量 Logistic 回归模型及其在 SPSS 中的实现。

10.4.1 Logistic 回归简介

在 Logistic 回归模型中，因变量是定性变量。首先从因变量是最简单的二分类变量的情况开始讨论。

二分类变量的情况十分普遍。比如在致癌因素的研究中，收集了若干人的健康记录，包括年龄、性别、抽烟史、日常饮食及家庭病史等变量的数据。因变量为一个人得了癌症 ($Y=1$)，或没有得癌症 ($Y=0$)。又比如在金融界，最关心的是企业的“健康”状况。自变量是公司的各项财务指标。而因变量是公司的偿付能力 (破产=0，有偿付能力=1)。这样的例子还有很多，可见二分类变量 Logistic 回归在实际问题中有着广泛的应用。

对于二分类变量的定性数据往往可以用 0 和 1 两个数值编码。现在要解决的问题是对因变量二值取一的概率建模而不是直接预测其取值。设有自变量 X_1 、 $X_2 \cdots X_n$ ，则令：

$$\pi = P(Y = 1 | X_1 = x_1, \dots, X_n = x_n) = \frac{e^{\beta_0 + \beta_1 x_1 + \dots + \beta_n x_n}}{1 + e^{\beta_0 + \beta_1 x_1 + \dots + \beta_n x_n}} \quad (10.8)$$

(10.8) 式称为 Logistic 回归函数，它关于参数 β_0 、 $\beta_1 \cdots \beta_n$ 是非线性的。然而可以通过 logit 变换将其线性化。这里不直接分析 π ，而是对其变换后的值进行分析。如果 π 是某事件发生的概率，那么比率 $\pi / 1 - \pi$ 则称为该事件的优势比。由于

$$1 - \pi = P(Y = 0 | X_1 = x_1, \dots, X_n = x_n) = \frac{1}{1 + e^{\beta_0 + \beta_1 x_1 + \dots + \beta_n x_n}} \quad (10.9)$$

则有

$$g(x_1, x_2, \dots, x_n) = \log\left(\frac{\pi}{1 - \pi}\right) = \beta_0 + \beta_1 x_1 + \dots + \beta_n x_n \quad (10.10)$$

优势比的对数称为 logit。由 (10.10) 式可知，logit 变换产生了参数 β_0 、 $\beta_1 \cdots \beta_n$ 的一个线性函数。

可见，拟合二分类变量的 Logistic 回归模型 (10.8) 的参数问题转换为拟合线性模型 (10.10) 的参数。通常采用极大似然法来估计参数。与最小二乘法不同的是，这里参数估计不存在精确解，只能通过迭代法获得极大似然估计的数值解。

与线性回归一样，拟合时也要考虑模型是否合适，哪些变量该保留，以及拟合效果等问题。线性回归常用的 R^2 、t 检验、F 检验等工具在这里都不再适用。在 10.4.3 节的例子中将看到 Logistic 回归的检验方法。

10.4.2 Binary Logistic过程的操作界面

首先介绍【Binary Logistic】过程的操作界面。执行【Analyze】/【Regression】/【Binary Logistic】命令，弹出如图 10-19 所示的对话框。

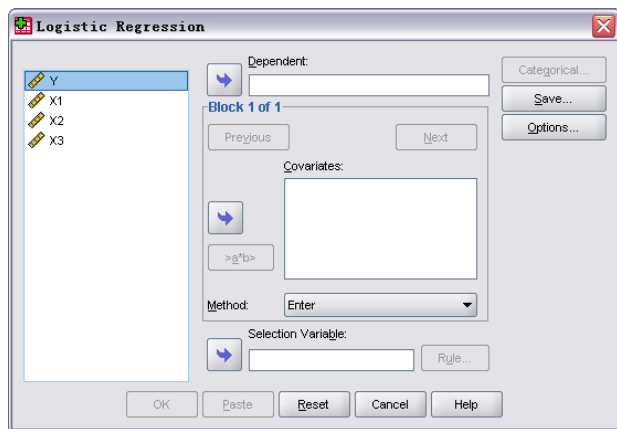


图 10-19 【Logistic】对话框

该对话框主要由以下几部分组成。


- 候选变量列表框

即左侧变量列表框。

- 【Dependent】框

选择 Logistic 回归的因变量，只能选择一个二值变量，否则最后的输出结果会出现警告。

- 【Covariates】框

选择 Logistic 回归的自变量。当候选变量框中同时选中两个或两个以上变量时，激活  按钮。该按钮用来将选中变量的交互作用项纳入【Covariates】框。

- 【Method】下拉列表

选择变量进入模型的方法。主要包括以下方法。

① Enter：强行进入法。

② Forward：Conditional/LR/Wald：依据条件参数似然比检验结果/偏似然比检验结果/Wald 检验结果剔除变量的向前剔除法。

③ Backward：Conditional/LR/Wald：依据条件参数似然比检验结果/偏似然比检验结果/Wald 检验结果剔除变量的向后剔除法。

- 【Selection Variable】框

选择筛选变量。当【Selection Variable】框选入一个变量后，则激活其后的【Rules】按钮。单击【Rules】按钮，弹出如图 10-20 所示的【Rules】对话框。该对话框主要用于给定变量的筛选条件。

- 【Categorical】

单击图 10-19 中的【Categorical】按钮，弹出如图 10-21 所示对话框。

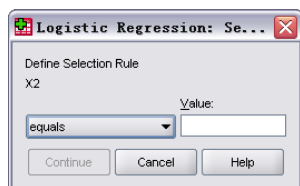


图 10-20 【Rules】对话框

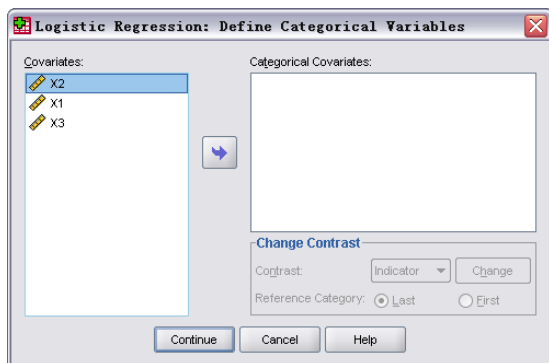


图 10-21 【Categorical】对话框

该对话框用于将某些数值型自变量定义为分类变量，主要由以下几部分组成。

① Covariates 框：候选变量列表框。

② Categorical Covariates 框：选择要将其定义为分类变量的变量。

③ Change Contrast 组：用于设置每个变量的哑变量组中的具体取值和对照组。其中 Contrast 下拉列表用于选择哑变量取值，Reference Category 单选框组用于设置选择第一个或最后一个水平为对照。一般而言，此项无需变动，直接采用系统默认值即可。

- 【Save】

单击图 10-19 中的【Save】按钮，弹出如图 10-22 所示的对话框。

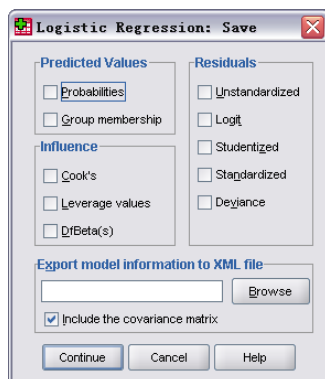


图 10-22 【Save】对话框

该对话框主要用于定义需要保存的中间统计量，由以下几部分组成。

① Predicted Values 复选框组：如表 10-27 所示，定义需要保存的预测值。

表 10-27 Predicted Values 复选框组各项具体作用

SPSS 中表示	作 用
Probabilities	保存每个观测量的预测概率值
Group membership	保存根据预测概率值判断观测值所属的类别

② Influence 复选框组：如表 10-28 所示，定义用于判断强影响点的统计量。

表 10-28 Influence 复选框组各项具体作用

SPSS 中表示	作 用
Cook's	保存删除当前记录后，模型残差会发生的变化量
Leverage values	保存杠杆值，即测量该数据点的影响强度
DfBeta	保存去掉该观察值后回归系数的变化值

③ Residuals：如表 10-29 所示，保存各类残差。

表 10-29 Residuals 复选框组各项具体作用

SPSS 中表示	作 用
Unstandardized	保存模型预测值对因变量观测值的原始残差
Logit	保存 Logit 残差
Studentized	保存学生化残差，即用 T 变换进行标准化后的残差
Standardized	保存用 U 变换进行标准化后的残差，此时均值为 0，标准差为 1
Deviance	保存 Deviance 残差

④ Export model information to XMLfile：选择是否将模型信息保存到 XML 文件中。

⑤ include the covariance matrix 复选框：选择是否保存变量间的相关矩阵。

• 【Options】

单击图 10-19 中的【Options】按钮，弹出如图 10-23 所示对话框。

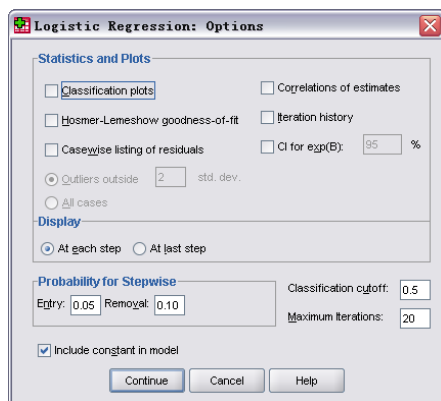


图 10-23 【Options】对话框

该对话框主要由以下几部分组成。

① Statistics and Plots 复选框组：如表 10-30 所示，定义一些重要的统计量和统计图形。

表 10-30 Statistics and Plots 复选框组各项具体作用

SPSS 中表示	作 用
Classification plots	绘制因变量实际分类和预测分类关系的图形
Hosmer-Lemeshow goodness-of-fit	计算 Hosmer-Lemeshow 拟合优度指标
Casewise listing of residuals	对于记录逐条列出或满足一定条件列出其残差和概率预测值、预测分类和实际分类
Correlations of estimates	计算参数估计值的相关系数矩阵
Iteration history	列出极大似然估计每一步的迭代估计值
CI for exp	计算参数值 95% 的置信区间

② Display 单选框组：定义是否输出模型迭代过程每一步的统计结果。

- At each step: 详细输出每一步的统计分析结果。
- At last step: 只输出最后一步的统计分析结果。

③ Probability for Stepwise 组：定义模型中变量进入或移出的概率值标准。

④ Classification cutoff 框：定义预测观测值分类的概率值大小。

⑤ Maximum iterations 框：定义最大迭代数。

⑥ Include constant in model 复选框：定义模型中是否包括常数项。

10.4.3 引例及结果解释

下面通过例子介绍【Binary Logistic】过程的操作及其结果。

例 10.4 诊断发现运营不良的金融商业机构是审计核查的一项重要功能。审计核查的分类失败会导致灾难性的后果。比如。美国 1980 年的储蓄一贷款的惨败事件。表 10-31 列出了 66 家公司的一些运营的财务比率，其中 33 家在 2 年后破产，另外 33 家在同期保持偿付能力。用变量 X_1 、 X_2 、 X_3 拟合一个 Logistic 回归模型。（数据来源：例解回归分析，中国统计出版社）

表 10-31 有偿付能力及破产公司的财务比率

序号	Y	X1	X2	X3	序号	Y	X1	X2	X3
1	0	-62.8	-89.5	1.7	34	1	43	16.4	1.3
2	0	3.3	-3.5	1.1	35	1	47	16	1.9
3	0	-120.8	-103.2	2.5	36	1	-3.3	4	2.7
4	0	-18.1	-28.8	1.1	37	1	35	20.8	1.9
5	0	-3.8	-50.6	0.9	38	1	46.7	12.6	0.9
6	0	-61.2	-56.2	1.7	39	1	20.8	12.5	2.4
7	0	-20.3	-17.4	1	40	1	33	23.6	1.5
8	0	-194.5	-25.8	0.5	41	1	26.1	10.4	2.1
9	0	20.8	-4.3	1	42	1	68.6	13.8	1.6
10	0	-106.1	-22.9	1.5	43	1	37.3	33.4	3.5
11	0	-39.4	-35.7	1.2	44	1	59	23.1	5.5
12	0	-164.1	-17.7	1.3	45	1	49.6	23.8	1.9
13	0	-308.9	-65.8	0.8	46	1	12.5	7	1.8
14	0	7.2	-22.6	2	47	1	37.3	34.1	1.5
15	0	-118.3	-34.2	1.5	48	1	35.3	4.2	0.9
16	0	-185.9	-280	6.7	49	1	49.5	25.1	2.6
17	0	-34.6	-19.4	3.4	50	1	18.1	13.5	4
18	0	-27.9	6.3	1.3	51	1	31.4	15.7	1.9
19	0	-48.2	6.8	1.6	52	1	21.5	-14.4	1
20	0	-49.2	-17.2	0.3	53	1	8.5	5.8	1.5
21	0	-19.2	-36.7	0.8	54	1	40.6	5.8	1.8
22	0	-18.1	-6.5	0.9	55	1	34.6	26.4	1.8
23	0	-98	-20.8	1.7	56	1	19.9	26.7	2.3
24	0	-129	-14.2	1.3	57	1	17.4	12.6	1.3
25	0	-4	-15.8	2.1	58	1	54.7	14.6	1.7
26	0	-8.7	-36.3	2.8	59	1	53.5	20.6	1.1
27	0	-59.2	-12.8	2.1	60	1	35.9	26.4	2
28	0	-13.1	-17.6	0.9	61	1	39.4	30.5	1.9
29	0	-38	1.6	1.2	62	1	53.1	7.1	1.9
30	0	-57.9	0.7	0.8	63	1	39.8	13.8	1.2
31	0	-8.8	-9.1	0.9	64	1	59.5	7	2
32	0	-64.7	-4	0.1	65	1	16.3	20.4	1
33	0	-11.4	4.8	0.9	66	1	21.7	-7.8	1.6

其中 X_1 、 X_2 、 X_3 、 Y 分别表示如下含义：

$$X_1 = \frac{\text{未分配利润}}{\text{总资产}}, \quad X_2 = \frac{\text{支付利息税金前的利润}}{\text{总资产}}, \quad X_3 = \frac{\text{销售额}}{\text{总资产}}$$

$$Y = \begin{cases} 0 & \text{两年后破产} \\ 1 & \text{两年后仍有偿付能力} \end{cases}$$

将表 10-31 中的数据保存在如图 10-24 所示的数据文件“gongsi.sav”中。

	Y	X1	X2	X3
1	.00	-62.80	-89.50	1.70
2	.00	3.30	-3.50	1.10
3	.00	-120.80	-103.20	2.50
4	.00	-18.10	-28.80	1.10

图 10-24 数据文件“gongsi.sav”的数据结构

执行以下操作：

执行【Analyze】/【Regression】/【Binary Logistic】命令，弹出【Logistic】对话框	
【Dependent】框：Y	定义 Y 为二分类因变量
【Covariates】框：X1、X2、X3	定义 X1、X2、X3 为自变量
单击【OK】按钮	定义完成

执行以上操作之后，生成表 10-32～表 10-40，分别解释如下。

表 10-32 是数据的基本信息，表中给出了数据进入模型的个数。

表 10-32 数据基本信息

Case Processing Summary			
Unweighted Cases ^a		N	Percent
Selected Cases	Included in Analysis	66	100.0
	Missing Cases	0	.0
	Total	66	100.0
Unselected Cases		0	.0
Total		66	100.0

a. If weight is in effect, see classification table for the total number of cases.

表 10-33 是因变量赋值表。在 SPSS 中，默认将二分类因变量中出现次数较多的值赋为 1。本例比较特殊，二分类变量的两种情况出现的次数正好一样。从表格中可以发现将“两年后破产”赋为 0，“两年后仍有偿付能力”赋为 1。

表 10-33 因变量赋值表

Dependent Variable Encoding	
Original Value	Internal Value
两年后破产 ^a	0
两年后仍有偿付能力	1

表 10-34 相当于模型的初始分类预测值。此时模型中不含任何自变量，只包括了常数项。表格左方代表实际观测值（Observed），右方代表模型的预测值（Predicted）和正确的预测率（Percentage Correct）。此时预测所有公司在两年后都仍有偿付能力。预测的正确率是 50%。

表 10-34 模型分类预测值

Classification Table ^{a,b}					
Observed			Predicted		
			Y		Percentage Correct
			两年后破产 ^c	两年后仍有 偿付能力	
Step 0	Y	两年后破产 ^c	0	33	.0
		两年后仍有偿付能力	0	33	100.0
Overall Percentage					50.0

a. Constant is included in the model.

b. The cut value is .500

表 10-35 是模型参数的检验结果。由于此时模型中只有常数项，Sig.取值为 1，模型没有任何统计学意义。

表 10-35 模型参数检验结果

Variables in the Equation						
	B	S.E.	Wald	df	Sig.	Exp(B)
Step 0 Constant	.000	.246	.000	1	1.000	1.000

表 10-36 是一个预分析的过程。即假设将未纳入模型的变量分别或一起纳入模型之后模型是否有统计学意义。从表格中 Sig.取值可知，除了单独纳入变量 X3 的模型没有统计学意义之外，其余模型都有显著的统计学意义。

表 10-36 未纳入模型的变量

Variables not in the Equation					
	Score	df	Sig.		
Step 1 Variables X1	31.621	1	.000		
0 X2	19.358	1	.000		
X3	2.800	1	.094		
Overall Statistics	37.613	3	.000		

表 10-37 是将 Block1 中变量纳入模型后模型的全局检验结果。共采用了 3 种检验方法，分别是步与步间的相对似然比检验（Step）、Block 间的相对似然比检验（Block）和模型间的相对似然比检验（Model）。由于本例中只有一个自变量组且采取的是强行进入法将所有变量纳入模型，所以 3 种检验方法的结果是一致的。模型有显著的统计学意义。

表 10-37 模型全局检验结果

Omnibus Tests of Model Coefficients				
	Chi-square	df	Sig.	
Step 1 Step	85.683	3	.000	
Block	85.683	3	.000	
Model	85.683	3	.000	

表 10-38 是模型摘要。主要给出了其-2 倍的似然比的对数值和两类决定系数。从数据上看，模型的拟合度不错。

表 10-38 模型情况摘要

Model Summary			
Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	5.813 ^a	.727	.969

a. Estimation terminated at iteration number 12 because parameter estimates changed by less than .001.

表 10-39 是将 Block1 中变量纳入模型后模型的分类预测值。此时模型预测的准确度就已经达到了 97%。

表 10-39 模型分类预测值

Classification Table ^a					
Observed			Predicted		
			Y		Percentage Correct
			两年后破产	两年后仍有偿付能力	
Step 1	Y	两年后破产	32	1	97.0
		两年后仍有偿付能力	1	32	97.0
Overall Percentage					97.0

a. The cut value is .500

表 10-40 是 Logistic 模型的拟合结果。表格从左到右依次表示变量及常数项的系数值 (B)，标准误 (S.E.)，Wald 卡方值，自由度 (df)，Sig.值，以及 Exp (B)。从 Wald 检验的 Sig.可知，各变量及常数项的系数都没有显著的统计学意义。

表 10-40 模型的参数拟合

Variables in the Equation							
		B	S.E.	Wald	df	Sig.	Exp(B)
Step	X1	.331	.301	1.213	1	.271	1.393
1 ^a	X2	.181	.107	2.862	1	.091	1.198
	X3	5.087	5.082	1.002	1	.317	161.979
	Constant	-10.153	10.840	.877	1	.349	.000

a. Variable (s) entered on step 1: X1, X2, X3.

10.4.4 小结

本节介绍了因变量为二分类变量的 Logistic 回归在 SPSS 中的实现。若因变量是多分类无序变量，那么可以调用【Multinomial】过程来实现其 Logistic 回归。若因变量是多分

类有序变量，则调用【Ordinal】过程。由于这两种 Logistic 回归在实际问题中不如二分类变量 Logistic 回归应用广泛。所以此处就不再单独介绍了。有兴趣的读者可以参阅 SPSS 的帮助文档。

10.5 非线性回归——Nonlinear过程

对于变量间的非线性关系，10.3 节引入了曲线拟合的方法。但是曲线拟合只能处理比较简单的模型。对于复杂的非线性关系，SPSS 又提供了 Nonlinear 过程。本节将通过例子介绍非线性回归的一般概念及其在 SPSS 中的实现。

10.5.1 非线性回归简介

在回归分析中，很多模型的回归参数是线性的。将这类模型称为线性回归模型，在 SPSS 中可以通过【Linear】过程实现。另外有一些特殊的模型，其回归参数不是线性的，但是可以通过转换变为线性的参数。这类模型又称为内蕴线性回归模型。SPSS 的【Curve Estimation】过程中的模型就是两变量间的内蕴线性回归模型。

但是仍然有一些模型，其回归参数不是线性的，也不能通过转换的方法将其变为线性的参数。将这类模型称为非线性回归模型。设因变量为 y ，自变量为 $x_1, x_2 \cdots x_m$ ，进行 n 次观测，则非线性回归模型可以写成 (10.11) 式所示的矩阵形式。

$$Y = f(X, \theta) + e \quad (10.11)$$

其中 Y 是一个 $n \times 1$ 的观测值向量， θ 是一个 $p \times 1$ 的回归参数向量， X 是一个 $n \times m$ 的自变量常数矩阵。 e 是一个 $n \times 1$ 的独立状态随机向量，服从多元正态分布。

一般地，用最小二乘估计法可以推导出非线性回归参数的估计值，从而使损失函数：

$$Q = e^T e = [Y - f(\theta)]^T [Y - f(\theta)] \quad (10.12)$$

为最小值。对 Q 中回归参数求偏导，有

$$\frac{\partial Q}{\partial \theta^T} = -2[Y - f(\theta)]^T \left[\frac{\partial f(\theta)}{\partial \theta^T} \right] = 0 \quad (10.13)$$

(10.13) 式的解即可作为回归参数的估计值。但是这组解是很难直接解出来的，只能通过迭代法求出其近似解。

SPSS 中的非线性回归是通过【Nonlinear】过程来实现的，它利用迭代的方法拟合各类复杂非线性模型的回归参数，有着十分强大的功能。并且【Nonlinear】过程还可以自定义各类损失函数。根据损失函数的不同，拟合的回归参数也不同。因此，【Nonlinear】过程最后拟合出来的回归参数也不一定是 (10.13) 式的解。

10.5.2 Nonlinear过程的操作界面

下面首先介绍【Nonlinear】过程的操作界面。执行【Analyze】/【Regression】/【Nonlinear】命令，弹出如图 10-25 所示对话框。

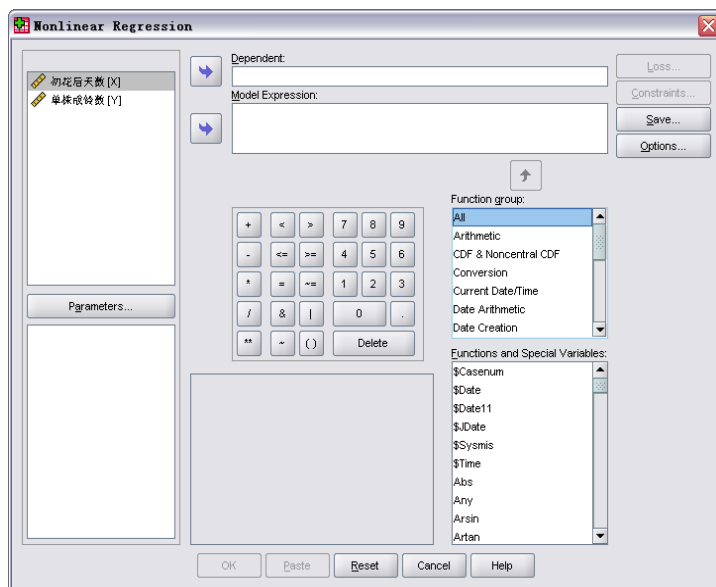


图 10-25 【Nonlinear】对话框

该对话框主要由以下几部分组成。

1. 候选变量列表框

即左上方的变量列表框。

2. 【Dependent】框

选择回归模型中的因变量。

3. 【Model Expression】框组

定义非线性回归模型的表达式。因为非线性模型实在是太多多种多样，所以 SPSS 中直接提供了软键盘和【Functions】组让用户自己定义非线性模型的表达式。在具体定义的时候，模型中的参数由用户直接通过键盘输入；模型中的变量由候选变量列表框选入；模型的运算符号由用户通过软键盘输入；模型的函数直接从【Functions】框中选入。比如图 10-25 中定义的模型表达式， a 、 b 、 c 表示模型中的参数， X 为自变量，EXP()为选入的指数函数。

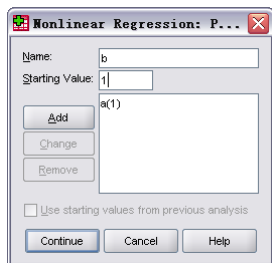


图 10-26 【Parameters】对话框

4. 【Functions】框组

包含【Function group】和【Functions and Special Variables】这两个列表框，几乎涵盖了所有常用的函数类型，其具体用法参见第 3 章。

5. 【Parameters】

单击【Parameters】按钮，弹出如图 10-26 所示的对话框。由 10.5.1 节可知，非线性回归模型的参数是通过迭代的

方法求得的。那么有必要指定迭代初始值，通常是通过 Parameters 对话框来定义的。该对话框包括以下几部分。

① Name 栏：选择模型中的参数，参数名必须与图 10-25 中的【Model Expression】框中的参数名一致。

② Starting Value：定义参数迭代初始值。

③ Add、Change、Remove 按钮组：添加、改变和移出定义的参数迭代初始值。

④ Use starting values from previous analysis 复选框：在连续使用非线性回归模型时，是否以上次模型的参数拟合值作为本次模型的迭代初始值。这样选择迭代初始值，可以大大减少模型的迭代次数。

6. 【Loss】

单击图 10-25 中的【Loss】按钮，弹出如图 10-27 所示的【Loss】对话框。

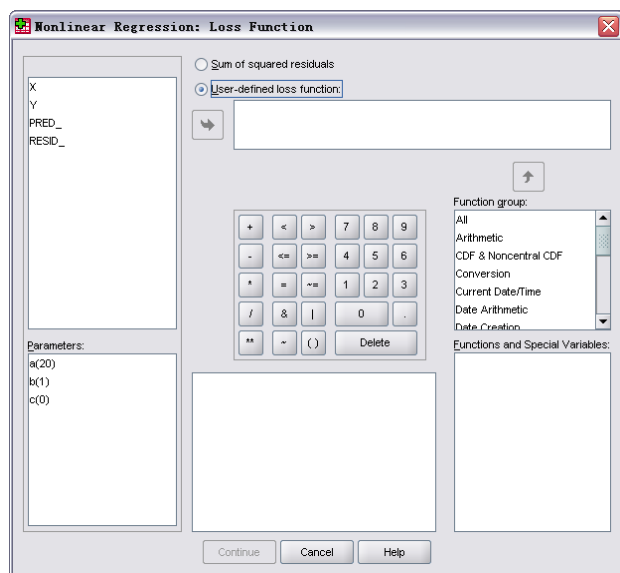


图 10-27 【Loss】对话框

该对话框主要用来定义回归模型的损失函数。包括以下两项。

- Sum of square residuals 单选框

选择该项则以 (10.12) 式所示的均方误差和作为损失函数。系统默认选项。

- User-defined loss function 单选框

用户自定义损失函数。选择该选项，才能激活 Loss 对话框的其余的各项。

① 损失函数定义框：即 User-defined loss function 单选框下方的列表框。

② 候选变量列表框：即左上方的列表框，列出了可以用来定义损失函数的变量。同时还包括“PRED_”和“RESID_”两项。其中“PRED_”代表变量的预测值，“RESID_”代表变量的残差。比如在损失函数定义框中选出“PRED_X”，则代表变量 X 的预测值进入损失函数模型。

① 候选参数列表框：即左下方列表框。

② 软键盘。

③ **Functions** 框组：可选函数列表框。

此时损失函数的定义其实是与图 10-25 中非线性回归模型的定义方法是类似的。根据用户自定义的损失函数，就可以求出在特定条件下非线性回归模型的参数估计值。

7. 【Constrains】

单击图 10-25 中的 **【Constrains】** 按钮，弹出如图 10-28 所示的 **【Constrains】** 对话框。

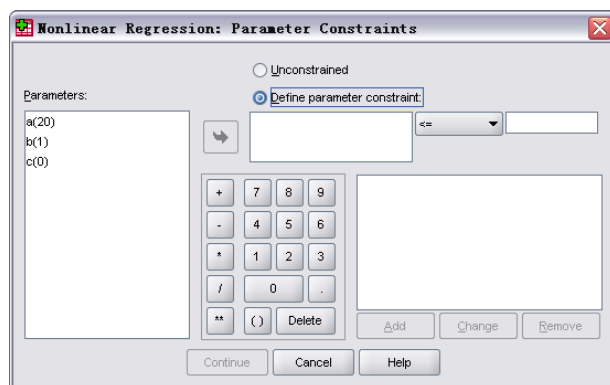


图 10-28 **【Constrains】** 对话框

该对话框主要用来定义模型中迭代参数的限制条件。包括以下两项。

① **Unconstrained** 单选框：系统默认选项，对参数不做任何限制。

② **Define parameter constraint** 单选框：用户自定义参数限制条件。其定义方法与前面介绍的非线性回归模型的方法类似。可以同时定义多组限制条件。

8. 【Save】

单击图 10-25 中的 **【Save】** 按钮，弹出如图 10-29 所示的 **【Save】** 对话框。

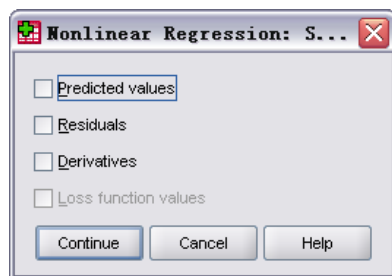


图 10-29 **【Save】** 对话框

该对话框主要用来定义需要保存的中间统计量。此时，将在原始的“*.sav”数据文件中新生成一列或几列变量用来保存这些统计量。主要包括预测值（Predicted values）、残差（Residuals）、各参数的导数（Derivatives）和损失函数值（Loss function values）。仅当用户自定义损失函数时，Loss function values 复选框才被激活。

9. 【Options】

单击图 10-25 中的【Options】按钮，弹出如图 10-30 所示的【Options】对话框。

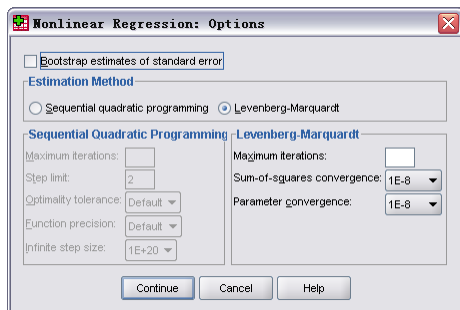


图 10-30 【Options】对话框

该对话框主要用来设置参数迭代拟合过程中的一些选项，具体包括以下几项。

- Bootstrap estimates of standard error 复选框

选择是否利用 Bootstrap 法估计参数的标准误差。

- Estimation Method 单选框组

定义参数的估计方法。

① Sequential quadratic programming 单选框：序列二次规划法，对于有无限制的模型都适用。

② Levenberg-Marquardt 单选框：Levenberg-Marquardt 法，只适用于无限制的模型。

- Sequential Quadratic Programming 组

进一步定义序列二次规划法的迭代过程。

① Maximum iterations 框：定义最大迭代次数，超出此次数则停止迭代。

② Step limit 框：定义迭代过程中步长允许的最大变化值。若超出此值，则停止迭代。

③ Optimality tolerance 下拉列表：最优容忍度，即定义模型中损失函数的精度。

④ Function precision 下拉列表：方程精度，即定义拟合的非线性回归模型的精度。

⑤ Infinite step size 框：定义迭代过程中所有参数允许的最大变化值。若超出此值，则停止迭代。

- Levenberg-Marquardt 组

进一步定义 Levenberg-Marquardt 法的迭代过程。

① Maximum iterations 框：定义最大迭代次数，超出此次数则停止迭代。

② Sum-of-squares convergence 下拉列表：定义迭代停止的条件。如果模型的损失函数的改变值小于此值时则停止迭代。

③ Parameter convergence 下拉列表：定义迭代停止条件。如果模型中所有参数的改变量小于此值时则停止迭代。

10.5.3 引例及结果解释

下面通过例子介绍【Nonlinear】过程的操作及其结果。

例 10.5 棉花单株在不同时期的成铃数（ Y ）与初花后天数（ X ）存在非线性的关系，假设这一非线性关系可用 Gompertz 模型表示

$$Y_i = \theta_1 e^{-\theta_2 e^{-\theta_3 X_i}} + e_i \quad (10.14)$$

某一棉花品种在 7 月 5 日至 9 月 3 日期间，每隔 5 天的单株成铃数观测值如表 10-41 所示。

表 10-41 原始观测数据值

X	5	10	15	20	25	30	35	40	45	50	55	60	65
Y	0.75	2.0	4.0	4.75	5.25	5.5	7.75	10.13	12.26	13.14	13.52	14.15	14.53

试根据观测值拟合模型中的参数。（数据来源：《线性模型分析原理》科学出版社）

对于例 10.5，执行以下操作：

STEP 01 建立数据文件“mianhua.sav”，其向量 X 代表初花后天数，向量 Y 代表棉花单株成铃数。

STEP 02 用非线性回归方法拟合模型（10.14）中的参数，执行以下操作：

执行【Analyze】/【Regression】/【Nonlinear】命令，弹出【Nonlinear】对话框	
【Dependent】框：Y	定义棉花单株成铃数为因变量
【Model Expression】框：a * EXP(-b * EXP(-c * X))	定义非线性回归模型表达式
单击【Parameters】按钮	弹出【Parameters】对话框
【Parameters】对话框：	
【Name】：a	定义参数 a 的初值
【Starting value】：20	将参数 a 的初值定义为 20
单击【Add】按钮	确认定义 $a = 20$
重复此过程定义 $b = 1$ ， $c = 0$	定义参数 b 、 c 迭代初值
单击【Continue】按钮	【Parameters】对话框定义完成
单击【OK】按钮	定义完成

以上操作将参数 θ_1 用 a 表示，定义其初始值为 20；参数 θ_2 用 b 表示，定义其初始值为 1；参数 θ_3 用 c 表示，定义其初始值为 0。执行以上操作之后，生成表 10-42～表 10-45，分别解释如下。

表 10-42 所示是模型参数拟合具体的迭代过程。在表中给出了模拟中参数在每一步迭代过程中的取值。分析表中注释可以发现，本例共迭代了 22 次，损失函数的差值小于 $1E-0.08$ ，从而停止迭代。

表 10-42 参数拟合的迭代过程

Iteration History ^b				
Iteration Number ^a	Residual Sum of Squares	Parameter		
		a	b	c
1.0	299.195	20.000	1.000	.000
1.1	7.7E+078	-1990.516	-98.474	.034

续表

Iteration Number ^a	Residual Sum of Squares	Parameter		
		a	b	c
1.2	6.8E+010	-190.516	-8.474	.034
1.3	289.332	8.068	1.455	.034
2.0	289.332	8.068	1.455	.034
2.1	42.687	14.929	1.919	.036
3.0	42.687	14.929	1.919	.036
3.1	14.326	20.592	3.127	.033
4.0	14.326	20.592	3.127	.033
4.1	10.541	17.212	3.410	.043
5.0	10.541	17.212	3.410	.043
5.1	6.832	19.111	3.464	.041
6.0	6.832	19.111	3.464	.041
6.1	6.764	18.985	3.490	.042
7.0	6.764	18.985	3.490	.042
7.1	6.763	19.034	3.486	.042
8.0	6.763	19.034	3.486	.042
8.1	6.763	19.026	3.487	.042
9.0	6.763	19.026	3.487	.042
9.1	6.763	19.028	3.486	.042
10.0	6.763	19.028	3.486	.042
10.1	6.763	19.027	3.486	.042

Derivatives are calculated numerically.

a. Major iteration number is displayed to the left of the decimal, and minor iteration number is to the right of the decimal.

b. Run stopped after 22 model evaluations and 10 derivative evaluations because the relative reduction between successive residual sums of squares is at most SSSCON=1.00E-008.

表 10-43 给出了模型中参数的估计值。具体给出了各参数的估计值 (Estimate)、标准误差 (Std. Error) 和 95% 置信区间 (95% Confidence Interval)。根据参数估计值, 可以得到例 10.5 的拟合结果为 $Y = 19.027e^{-3.486e^{-0.042x}}$ 。

表 10-43 参数估计值

Parameter Estimates				
Parameter	Estimate	Std. Error	95% Confidence Interval	
			Lower Bound	Upper Bound
a	19.027	2.521	13.410	24.645
b	3.486	.449	2.487	4.486
c	.042	.008	.023	.061

表 10-44 是各参数的相关系数矩阵。从表中数据可以发现，各参数之间的相关系数还是比较大的。

表 10-44 参数相关系数矩阵

Correlations of Parameter Estimates

	a	b	c
a	1.000	-.611	-.945
b	-.611	1.000	.815
c	-.945	.815	1.000

表 10-45 是模型的显著性检验结果，采用的是方差分析的方法。此时决定系数为 0.977，说明模型的拟合效果很好。

表 10-45 方差分析表

ANOVA^a

Source	Sum of Squares	df	Mean Squares
Regression	1173.954	3	391.318
Residual	6.763	10	.676
Uncorrected Total	1180.718	13	
Corrected Total	287.968	12	

Dependent Variable: 单株成铃数

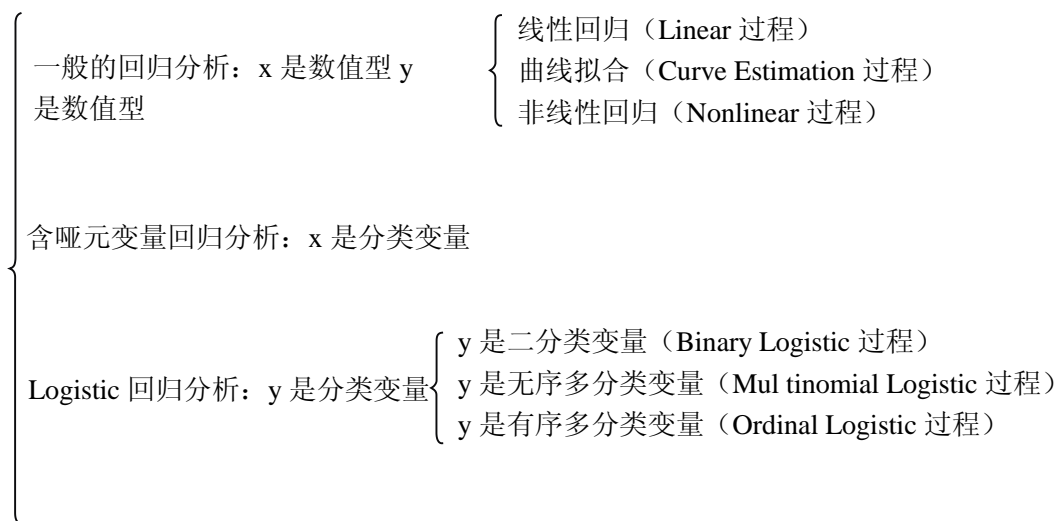
a. R squared=1-(Residual Sum of Squares)/(Corrected Sum of Squares)= .977.

10.5.4 小结

本节所介绍的【Nonlinear】过程是处理非线性回归问题的常用的一种方法，其功能十分强大。但是在拟合的过程中，是由用户自定义回归模型。可见，如何定义恰当的模型是解决问题的关键。这既依赖于模型中的数据特征，更依赖于模型中问题的实际背景。所以本书反复强调，在用 SPSS 软件解决问题的时候一定不能脱离问题的实际背景及其统计意义。

10.6 本章小结

本章介绍了回归分析在 SPSS 中的实现。按照自变量 x 和因变量 y 的类型，可将回归分析分成如下几类：



其中, 本章详细介绍了一般回归分析和二分类变量的 Logistic 回归分析。对于自变量 x 是分类变量的回归分析, 应首先利用【Transform】菜单下的【Compute Variable】过程将其设置为哑元变量, 再用一般的回归分析方法来处理。具体的过程请读者自己研究学习。

第 11 章 聚类分析与判别分析

聚类分析和判别分析都是用于解决分类问题的多元统计分析方法。SPSS 的【Classify】子菜单提供了多种聚类分析和判别分析方法。本章将通过实际例子来学习聚类分析和判别分析的一般步骤及其在 SPSS 中的实现过程。本章内容包括：

- 聚类分析与判别分析相关原理简介
- K-均值聚类分析——K-means Cluster 过程
- 系统聚类法——Hierarchical Cluster 过程
- 两步聚类法——TwoStep Cluster 过程
- 判别分析——Discriminant 过程

11.1 聚类分析与判别分析相关原理简介

本节主要介绍聚类分析与判别分析的意义、重要性和类型。对于各类具体的聚类分析和判别分析方法，将在本章后续几节给出详细地介绍。

11.1.1 聚类分析

聚类分析又称群分析，是研究（样品或指标）分类问题的一种多元统计方法，所谓类，通俗地说就是指相似元素的集合。在实际问题中存在大量的分类问题，比如，要对某些大城市的物价指数进行考察，而物价指数很多，有农用生产物价指数、服务项目物价指数、食品消费物价指数，等等。由于要考虑的物价指数很多，通常需要先对这些物价指数进行分类，这就要用到聚类分析了。总体说来，聚类分析就是把没有分类信息的资料按照相似程度归类。

聚类分析的内容非常丰富，大致可以分为两类：系统聚类法和非系统聚类法。其中系统聚类法是应用最广泛的一种方法。

11.1.2 判别分析

判别分析是判别样品所属类型的一种统计方法。与聚类分析一样，判别分析也是用于解决分类问题的，不同之处在于，判别分析是在已知研究对象分成若干类型（或组别）并已取得各种类型的一批已知样品的观测数据的基础上，根据某些准则建立判别式，然后对未知类型的样品进行判别分析。

判别分析同样内容丰富，方法很多，按照判别准则可以分为距离判别、Bayes（贝叶斯）判别和 Fisher（费歇）判别等。

由于聚类分析和判别分析都是处理分类问题，因此 SPSS 将它们放在同一个子菜单【Classify】中。该菜单如图 11-1 所示，主要包括 3 大模块，从上到下依次是聚类分析、判别分析以及最近邻域分析。本章将具体介绍聚类分析和判别分析这两大模块中的几个重要过程。

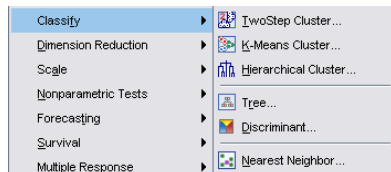


图 11-1 【Classify】子菜单

11.2 K-均值聚类分析——K-means Cluster过程

K-均值聚类法又被称为快速聚类法，它是非系统聚类法中最常用的聚类法，它的优点是占有内存少、计算量小、处理速度快，特别适合大样本的聚类分析。

11.2.1 K-均值聚类法基本原理

设待分类的观测量集为 $\{x_1, x_2, \dots, x_N\}$ ， N 为观测量数目， x_{ij} 表示第 i 个观测量的第 j 个变量值。给定聚类数目 n 和聚类比例系数 θ 。

K-均值聚类法的具体步骤如下。

(1) 根据给定的聚类数目 n ，按照一定的准则选择某些观测量 $\{z_1, z_2, \dots, z_n\}$ 作为初始聚类中心，简称聚心。

(2) 计算每个观测量到各个聚心的欧式距离，即令 $d_{ij} = \|x_i - z_j\| = [\sum_{k=1}^n (x_{ik} - z_{jk})^2]^{\frac{1}{2}}$ 。

按就近原则将每个观测量选入一个类中，然后计算各个类中的中心位置，即均值，作为新的聚心。

(3) 使用计算出来的新聚心重新进行分类，分类完毕后继续计算各类的中心位置，作为新的聚心，如此反复操作，直到两次迭代计算的聚心之间距离的最大改变量小于初始聚心间最小距离的 θ 倍时，或者达到迭代次数的上限时，停止迭代。

K-均值聚类法的缺点是应用范围有限，因为它要求用户指定分类数目，只能对观测量聚类，而不能对变量聚类，且所使用的聚类变量必须都是连续性变量。

11.2.2 K-means Cluster过程界面操作介绍

执行【Analyze】/【Classify】/【K-means Cluster】命令，弹出如图 11-2 所示的【K-Means Cluster Analysis】（K-均值聚类分析）对话框，下面介绍其中的元素。

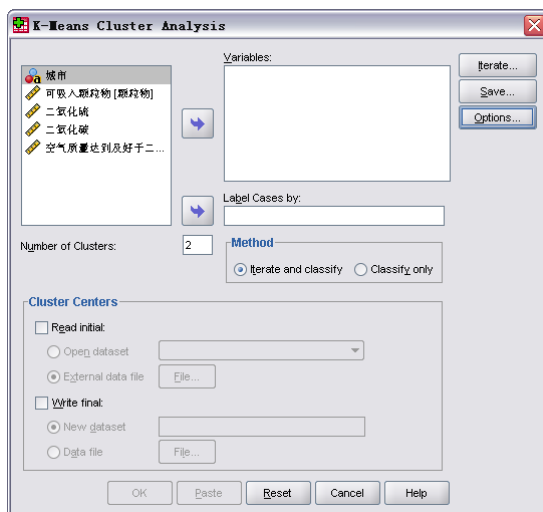


图 11-2 K-均值聚类分析对话框

1. 【Variables】框

变量框，放置用于进行 K-均值聚类的变量。

注意 考虑到各个变量的单位可能不同，因此，有必要对各数值型变量做标准化处理，即计算各变量经过 Z 变换后的取值，并建立新变量保存这些取值，详见 6.3 节。

2. 【Label Cases by】框

标记观测量框，用于放置标记变量，该变量的作用相当于观测量记录号的作用。

3. 【Number of Clusters】栏

类数目栏，设置聚类数，系统默认为 2。

4. 【Method】框

方法栏，用于选择聚类方法。系统默认选项是“Iterate and classify”，该选项是指在迭代过程中不断地更新聚类中心；“Classify only”选项是指迭代过程中聚类中心一直不变。

5. 【Cluster Centers】框

聚心框，用于设置最终聚心和初始聚心存取。如果不选第一项，系统将自动生成初始聚心；如果选择第二项，则将最终聚心保存到指定的文件或数据集中。

6. 【Iterate】子对话框

迭代子对话框，用于设置迭代参数。单击【Iterate】按钮，弹出此对话框，如图 11-3 所示。

① Maximum Iterations 栏：最大迭代次数栏，栏内输入迭代次数的上限，系统默认为 10。

② Convergence Criterion 栏：收敛标准值栏，栏内输入一个不超过 1 的正数，表示比例系数 θ 的取值。

③ Use running means 选项：使用移动平均选项。选择此项，则表示在迭代过程中每分配一个观测量到某类后就立刻计算新的聚心；不选此项，则表示当所有观测量分配完以后再计算各类聚心。

7. 【Save New Variables】子对话框

保存新变量子对话框，用于选择保存新变量的方式。单击【Save】按钮，弹出此对话框，如图 11-4 所示。

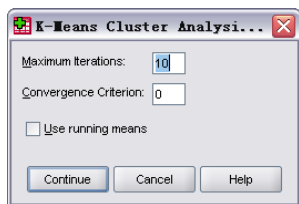


图 11-3 【Iterate】子对话框

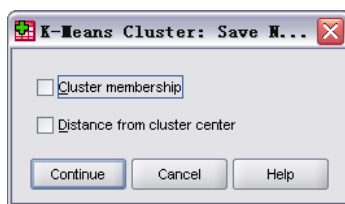


图 11-4 【Save New Variables】子对话框

① Cluster membership 选项：聚类成员选项。若选择此项，则工作文件中将建立一个名为“QCL_1”的变量，其值为各观测量的聚类后最终所属的类别。

② Distance from cluster center 选项：聚类中心距离选项。若选择此项，工作文件中将建立一个名为“QCL_2”的变量，其值为各观测量与所属类的聚心之间的欧氏距离。

8. 【Options】子对话框

选项子对话框，其中包括两个框。单击【Options】按钮，弹出此对话框，如图 11-5 所示。

• Statistics 框

统计框，用于指定输出统计量值。

① Initial cluster centers 选项：初始聚心选项，输出初始聚心，系统默认选项。

② ANOVA table 选项：方差分析表选项，输出方差分析表。在聚类的过程中，可能引入了无关变量，这样会降低聚类的效果，可见，在使用方差分析表来分析变量在类间的差异时，若发现差异很小的变量，就可以将它从【Variables】框中去除。

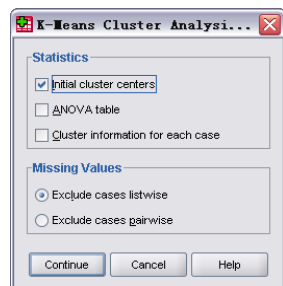


图 11-5 【Options】子对话框

注意 做聚类分析时，并不是变量越多越好，无关变量的存在可能会影响分类的准确性。

③ Cluster information for each case 选项：每个观测量的聚类信息选项，显示每个观测量最终被聚入的类别、各观测量与最终聚心的欧氏距离，以及最终各类聚心之间的欧氏距离。

• Missing Values 框

缺失值框，用于指定缺失值的处理方式。

11.2.3 引例及结果解释

下面通过一个例子介绍【K-means Cluster】过程的操作及其结果。

例 11.1 以数据文件“主要城市空气质量指标.sav”为例，其中的变量包括“城市”、显示该城市空气质量的指标变量“颗粒物”、“二氧化硫”、“二氧化碳”和“天数”，其中“天数”是指空气质量达到或好于二级的天数。数据如表 11-1 所示。（数据来源：《中国统计年鉴》—2005）

表 11-1 主要城市空气质量指标表

城市	颗粒物	二氧化硫	二氧化碳	天数	城市	颗粒物	二氧化硫	二氧化碳	天数
北京	0.149	0.055	0.071	229	武汉	0.130	0.048	0.054	247
天津	0.111	0.073	0.052	299	长沙	0.140	0.084	0.033	219
石家庄	0.123	0.087	0.042	279	广州	0.099	0.077	0.073	304
太原	0.175	0.087	0.022	224	南宁	0.078	0.061	0.034	348
呼和浩特	0.080	0.045	0.038	311	海口	0.033	0.007	0.013	366
沈阳	0.137	0.052	0.035	301	重庆	0.142	0.113	0.067	243
长春	0.085	0.013	0.032	345	成都	0.115	0.067	0.048	309
哈尔滨	0.113	0.042	0.060	298	贵阳	0.083	0.094	0.024	337
上海	0.099	0.055	0.062	311	昆明	0.085	0.069	0.040	351
南京	0.121	0.045	0.055	295	拉萨	0.052	0.003	0.020	358
杭州	0.110	0.049	0.055	292	西安	0.142	0.049	0.033	260
合肥	0.110	0.013	0.017	313	兰州	0.172	0.071	0.045	204
福州	0.074	0.010	0.041	358	西宁	0.127	0.024	0.027	280
南昌	0.099	0.057	0.029	330	银川	0.122	0.054	0.040	323
济南	0.149	0.045	0.038	210	乌鲁	0.114	0.102		258
郑州	0.111	0.057	0.037	298	木齐				

利用快速聚类过程将文件中的 31 个城市按空气质量分成 5 类。执行以下操作：

执行【Analyze】/【Descriptives Statistics】/【Descriptives】命令，弹出【Descriptives】对话框

Variables：颗粒物、二氧化硫、二氧化碳、天数

勾选 Save standardized values as variables 项 对聚类变量做标准化处理

单击【OK】按钮

执行【Analyze】/【Classify】/【K-means Cluster】命令，弹出对话框

Variables：Z 颗粒物、Z 二氧化硫、Z 二氧化碳、Z 天数 将聚类分析变量移入【Variables】框

Label Cases by：城市

以变量“城市”作为标记变量

Number of Clusters：5

指定聚类数为 5

单击【Iterate】按钮

弹出【Iterate】子对话框

Convergence Criterion：0.02

指定收敛标准值为 0.02

单击【Continue】按钮

回到主对话框

单击【Save】按钮	弹出【Save New Variables】子对话框
勾选其中两个选项	选择在文件中添加两个新变量
单击【Continue】按钮	回到主对话框
单击【Options】按钮	弹出【Options】子对话框
选择 Exclude cases Pairwise	定义缺失值处理方式
单击【Continue】按钮	回到主对话框
单击【OK】按钮	生成以下结果

表 11-2 是初始类聚心表，表中列出了由系统给出的各类的初始聚类中心。

表 11-2 初始类聚心表

	Cluster				
	1	2	3	4	5
Zscore: 可吸入颗粒物	1.14931	-.91521	.46114	1.96261	-2.47924
Zscore (二氧化硫)	-.00345	1.38844	-1.10983	1.13861	-1.71655
Zscore (二氧化碳)	1.88010	-1.08848	-.89900	-1.21480	-1.78325
Zscore: 空气质量达到 及好于二级的天数	-1.38581	.93288	-.29087	-1.49316	1.55549

表 11-3 是迭代史表，表中列出了迭代过程中各类聚心的改变值。表下的注释指出聚类过程经过 6 次迭代才终止，初始聚类中心之间的最小距离为 2.976。

表 11-3 迭代史表

Iteration	Change in Cluster Centers				
	1	2	3	4	5
1	1.364	1.340	.958	.867	.973
2	.181	.224	.310	.344	.294
3	.096	.238	.215	.000	.000
4	.139	.250	.000	.000	.000
5	.086	.242	.000	.000	.000
6	.000	.000	.000	.000	.000

a. Convergence achieved due to no or small change in cluster centers. The maximum absolute coordinate change for any center is .000. The Current iteration is 6. The minimum distance between initial centers is 2.976.

表 11-4 是最终聚类中心表，表中列出了各类的最终聚类中心。

表 11-4 最终聚类中心表

	Cluster				
	1	2	3	4	5
Zs core: 可吸入颗粒物	.20568	-.85265	.39336	1.46212	-1.60338
Zs core (二氧化硫)	.45159	.36058	-.48526	.59434	-1.67194
Zs core (二氧化碳)	1.06475	-.52003	-.61477	-.42529	-.93058
Zs core: 空气质量达到 及好于二级的天数	-.28372	.89852	.04906	-1.70248	1.35689

表 11-5 是每类中的样品数目表，表中列出了每个类中的观测量数目，以及有效观测量数目和缺失观测量数目。

表 11-5 每类中的样品数目表

Number of Cases in each Cluster		
Cluster	1	12.000
	2	5.000
	3	6.000
	4	4.000
	5	4.000
Valid		31.000
Missing		.000

11.3 系统聚类法——Hierarchical Cluster过程

系统聚类法是效果最好且最常用方法之一，本节将具体介绍系统聚类法以及它在 SPSS 中的应用。

11.3.1 系统聚类法基本原理

系统聚类法的基本思想是：首先视 n 个观测量（或变量）各自为一类，然后找性质最接近的两个类合并成一个新类，计算在新的类别划分下计算各类之间的距离，再将性质最接近的两类合并，直到所有观测量（或变量）聚成一类为止。

围绕着这个基本思想，系统聚类法具体到算法上，需要定义两个类之间距离测量的方法，定义距离测度方法等等，SPSS 为系统聚类法提供了 7 种类间距离测量的方法，以及针对不同类型变量的多种距离算法。

系统聚类法最大的优点在于既可以对观测量也可以对变量进行聚类，所使用的变量既可以是连续变量也可以是分类变量，提供的距离计算方法和结果显示方法也很丰富。

11.3.2 Hierarchical Cluster过程界面操作介绍

执行【Analyze】/【Classify】/【Hierarchical Cluster】命令，弹出如图 11-6 所示的【Hierarchical Cluster Analysis】（分层聚类分析）对话框，下面介绍其中元素。

1. 【Variables】框

变量框，放置用于进行分层聚类的变量。

2. 【Label Cases by】栏

标记观测量框。只有对观测量进行聚类时此项才被激活。

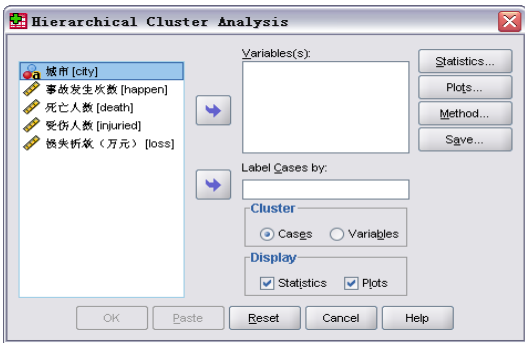


图 11-6 分层聚类分析对话框

3. 【Cluster】单选框

聚类单选框，用于选择聚类的内容是观测量（Cases）还是变量（Variables）。

4. 【Display】复选框

显示复选框，选择显示的内容，包括显示统计量值（Statistics）和显示图形（Plots）。

5. 【Statistics】子对话框

统计量子对话框，用于选择要输出的统计量。单击【Statistics】按钮，弹出此对话框，如图 11-7 所示。只有在主对话框中勾选了“Statistics”选项时此项才被激活。其中包括 3 种统计量值。

① Agglomeration schedule: 输出一张概述聚类进度的表格，系统默认选项。

② Proximity matrix: 输出一个相似性矩阵来显示各项间的距离。

③ Cluster Membership 单选框：样品隶属类单选框。有 3 个选项，分别表示不输出样品隶属类表（None），此项为系统默认的；指定一个分类数目，然后输出样品隶属表（Single solutions）；指定两个分类数 $m < n$ ，输出分类数从 m 到 n 的各种分类的样品隶属表。

6. 【Plots】子对话框

图形子对话框，用于选择要输出的图形。单击【Plots】按钮，弹出此对话框，如图 11-8 所示。只有在主对话框中勾选了“Plots”选项时此项才被激活。

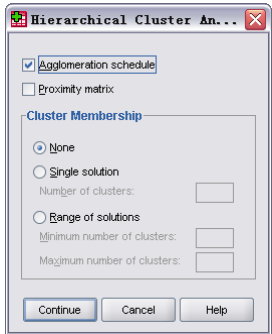


图 11-7 【Statistics】子对话框

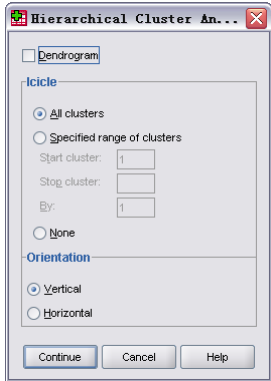


图 11-8 【Plots】子对话框

其中包括两种图形。

① **Dendrogram**: 龙骨图。将在结果分析中介绍。

② **Icicle**: 冰柱图。其中有三个选项，第一项是指显示全部聚类结果的冰柱图，为系统默认选项；第二项是指限制聚类解范围，表示从最小聚类解（**Start** 后面的数）开始，以 **By** 后面的数为步长，到最大聚类解（**Stop** 后面的数）为止。第三项是指不输出冰柱图。下面有一个 **Orientation** 单选框，用于选择冰柱图是垂直的（**Vertical**）还是水平的（**Horizontal**）。

7. 【Method】子对话框

方法子对话框，用于选择具体的聚类方法。单击【**Method**】按钮，弹出此子对话框。如图 11-9 所示，此对话框共包含 4 个部分。

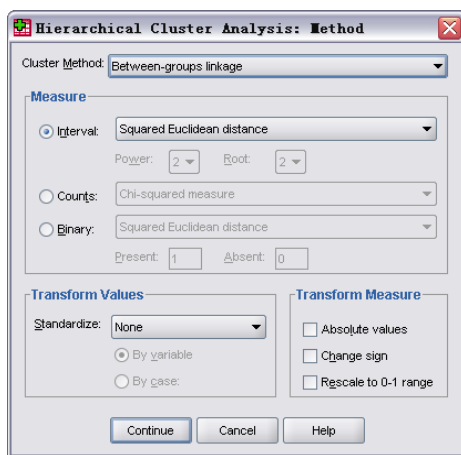


图 11-9 【Method】子对话框

• Cluster Method 下拉列表

聚类方法下拉列表，用于选择不同类间距离的测量方法，SPSS 提供了 7 种方法，如表 11-6 所示。

表 11-6 不同类间距离的测量方法列表

聚类方法	中 文	意 义
Between-groups linkage	组间连接法	合并两类使得两类间的平均距离最小，系统默认选项
Within-groups linkage	组内连接法	合并两类使得合并后的类中所有项间的平均距离最小
Nearest Neighbor	最近距离法	定义类与类之间的距离为两类中最近的样品之间的距离
Furthest Neighbor	最远距离法	定义类与类之间的距离为两类中最远的样品之间的距离
Centroid Clustering	重心法	定义类与类之间的距离为两类中各样品的重心之间的距离
Median Clustering	中位数法	定义类与类之间的距离为两类中各样品的中位数之间的距离
Ward's Method	Ward 最小偏差平方和法	聚类中使类内各样品的偏差平方和最小，类间偏差平方和尽可能大

注意 选择不同的测量方法，所得到的分类结果可能会大相径庭。

• Measure 单选框

测度单选框，用于选择距离测度方法下面的 3 个选项，分别为不同类型的变量所提供。其中算法的具体意义见 9.4 节。

① **Interval**：为间隔测度的连续型变量提供距离算法，在下拉菜单中列出了所提供的算法名称，其中系统默认选项是 **Squared Euclidean distance**（欧氏距离的平方）。

② **Count**：为频数计数变量提供测度计数数据的不相似性方法，在下拉菜单中列出了所提供的方法名称，其中系统默认选项是 **Chi-squared measure**（卡方测度）。

③ **Binary**：为二元变量提供二值数据的不相似性测度，在下拉菜单中列出了所提供的方法名称，其中系统默认选项是 **Squared Euclidean distance**（二元变量欧氏距离的平方）。

• Transform Values 框

转换值框，用于选择数据标准化方法。其中的 **Standardize** 下拉列表用于选择对变量或者对观测量的数据标准化方法。下面有两个单选项，分别是 **By variable**（对变量实行标准化）和 **By case**（对观测量实行标准化）。

• Transform Measures 复选框

转换测度复选框，用于选择测度转换方法。系统提供了三种方法，包括 **Absolute Values**（绝对值转换法）、**Change sign**（变号转换法）和 **Rescale to 0-1 range**（重新调节测度值到范围 0-1 转换法）。

8. 【Save New Variables】子对话框

保存新变量子对话框。单击【**Save**】按钮，弹出此子对话框，如图 11-10 所示。只有对观测量进行聚类时此项才被激活。其中只有一个单选框，与图 11-7 中的单选框相似，每个选项的意义也相同，只是这里不是输出表格，而是建立新变量来存储结果。

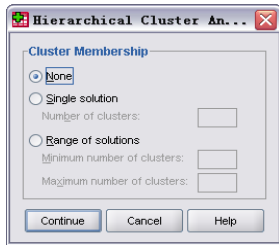


图 11-10 【Save New Variables】子对话框

11.3.3 引例及结果解释

下面通过两个例子来分别介绍【**Hierarchical Cluster**】过程对观测量和变量的聚类操作及其结果。

例 11.2 以数据文件“各地区交通事故情况(2004).sav”为例，其中的变量包括“city”——城市名称、“happen”——事故发生次数、“death”——死亡人数、“injured”——受伤人数、“loss”——损失折款（万元），数据如表 11-7 所示。（数据来源：《中国统计年鉴》—2005）。

表 11-7 各地区交通事故情况表（2004）

City	happen	death	injured	loss	City	happen	death	injured	loss
北京	8536	1631	8284	4058.0	湖北	13584	2530	11724	7741.2
天津	5485	992	4859	3273.8	湖南	16116	3824	18382	8546.7
河北	15095	4565	15088	7760.1	广东	68423	10657	78562	24127.2
山西	17206	4172	13843	8403.2	广西	13263	3641	14258	5427.0
内蒙古	9889	2239	8500	2798.8	海南	2041	520	2232	819.4
辽宁	12985	3346	10212	7398.9	重庆	11109	1619	12619	2894.0

续表

City	happen	death	injured	loss	City	happen	death	injured	loss
吉林	9955	2515	6892	5046.3	四川	28484	4890	28218	11253.8
黑龙江	8532	2463	8318	4331.4	贵州	3395	1831	3488	2353.3
上海	27136	1543	11304	19148.6	云南	11421	3210	7250	5195.9
江苏	31431	8100	23493	17655.8	西藏	1097	594	1001	956.4
浙江	50039	7549	50748	27853.4	陕西	13348	2949	8988	6242.6
安徽	18006	4794	17880	6996.0	甘肃	6361	1992	5566	2512.2
福建	24274	4393	24314	9408.8	青海	1212	742	1428	565.4
江西	10531	2760	10004	6370.0	宁夏	4216	864	3802	1312.0
山东	39815	7804	36077	13283.3	新疆	8364	2881	8902	2964.8
河南	26540	5467	24628	1244.3					

利用系统聚类法对各地区按照其交通情况进行聚类，执行以下操作：

执行【Analyze】/【Descriptives Statistics】/【Descriptives】命令，弹出【Descriptives】对话框

Variables : happen、death、injured、loss

勾选 Save standardized values as variables

对聚类变量做标准化处理

单击【OK】按钮

执行【Analyze】/【Classify】/【Hierarchical Cluster】命令，弹出对话框

Variables : Zhappen、Zdeath、
Zinjured、Zloss

将聚类分析变量移入【Variables】框

单击【Plots】按钮

弹出【Plots】子对话框

勾选 Dendrogram

输出龙骨图

Icicle : None

不输出冰柱图

单击【Continue】按钮

回到主对话框

单击【OK】按钮

生成以下结果

表 11-8 所示是观测量概述表。

表 11-8 观测量概述表

Case Processing Summary^{a,b}

Cases					
Valid		Missing		Total	
N	Percent	N	Percent	N	Percent
31	100.0	0	.0	31	100.0

a. Squared Euclidean Distance used

b. Average Linkage (Between Groups)

表 11-9 所示是聚类进度表。这个表格主要用来描述系统聚类法的具体实现步骤。表中第一列 Stage 代表聚类的步数，第二列 Cluster Combined 代表该步具体合并的是哪两类，第三列 Coefficients 代表类与类之间的距离测度系数，第四列 Stage Cluster First Appears 代表该步聚类合并的两类的上一次出现的步骤数，最后一列 Next 代表本步生成的新类下一次合并将出现在第几步。

以 Stage 为 1 时,即以第一步聚类为例,将观测量 26 (Cluster 1) 和观测量 29 (Cluster 2) 聚类合并,其中的距离测度系数 (Coefficients) 是 0.008,这个类的下一次聚类合并 (Next Stage) 将出现在步骤 2。其中 Stage Cluster First Appears (复聚类首次出现的步骤数) 下的 Cluster 1 和 Cluster 2 为 0 时表示非类的合并,若不为 0 则表示某个类,且显示的数字就是生成该类的那个步骤数。

表 11-9 聚类进度表

Agglomeration Schedule

Stage	Cluster Combined		Coefficients	Stage Cluster First Appears		Next Stage
	Cluster 1	Cluster 2		Cluster 1	Cluster 2	
1	26	29	.008	0	0	2
2	21	26	.013	0	1	7
3	7	8	.029	0	0	10
4	14	27	.046	0	0	8
5	3	4	.062	0	0	12
6	24	28	.062	0	0	17
7	21	30	.075	2	0	21
8	14	25	.081	4	0	11
9	5	31	.082	0	0	10
10	5	7	.106	9	3	18
11	6	14	.109	0	8	15
12	3	18	.131	5	0	16
13	1	22	.134	0	0	18
14	16	23	.156	0	0	20
15	6	17	.165	11	0	19
16	3	12	.166	12	0	23
17	2	24	.176	0	6	21
18	1	5	.229	13	10	22
19	6	20	.245	15	0	22
20	13	16	.340	0	14	26
21	2	21	.343	17	7	25
22	1	6	.474	18	19	23
23	1	3	1.190	22	16	25
24	10	15	1.381	0	0	26
25	1	2	2.075	23	21	29
26	10	13	2.880	24	20	27
27	9	10	6.289	0	26	29
28	1	9	6.774	26	0	30
29	10	19	13.817	27	0	30
30	1	10	26.357	28	29	0

图 11-11 所示为龙骨图，图中横向距离表示差异的大小，从图中可以清晰地看出整个的观测量的聚类过程。

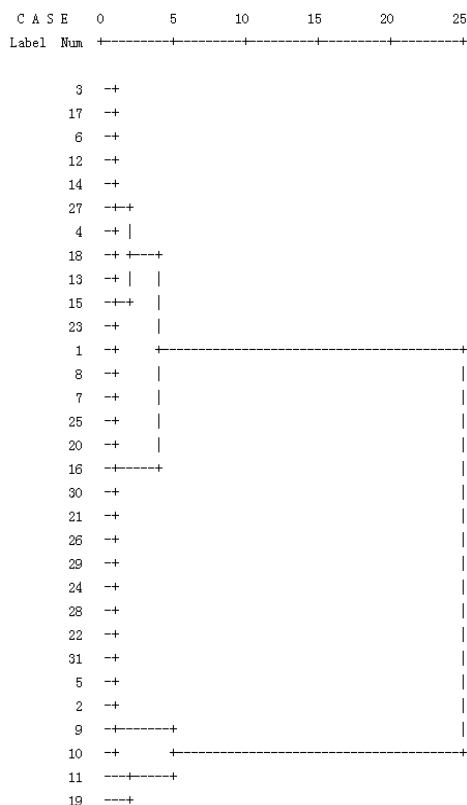


图 11-11 龙骨图

下面的这个例子主要是介绍变量的聚类过程。

例 11.3 以数据文件“主要城市日照时数.sav”为例，其中的变量包括城市的名称“city”、各个月份的日照数“Jan”、“Feb”、…、“Dec”。数据如表 11-10 所示。（数据来源：《中国统计年鉴》—2005）。

表 11-10 主要城市日照时数表

city	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
北京	194.7	213.5	243.6	248.2	253.3	202.0	203.2	187.4	198.9	225.2	201.4	144.0
天津	161.7	185.2	166.8	214.3	221.0	182.5	179.5	149.8	178.7	194.7	172.8	119.1
石家庄	193.8	219.2	220.9	240.9	277.9	213.4	185.4	152.1	203.4	220.7	197.5	97.9
太原	138.9	221.4	203.0	266.7	262.5	195.8	214.7	165.1	200.4	220.7	207.4	123.2
呼和浩特	187.5	207.8	237.6	284.9	302.2	234.5	291.0	227.4	240.3	223.3	185.9	135.6
沈阳	165.4	180.7	231.7	245.3	219.3	230.3	133.0	198.3	211.1	229.9	132.2	114.5
大连	163.5	195.3	223.1	276.9	243.4	190.0	228.5	174.0	202.7	228.4	172.9	167.0

续表

city	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
长春	194.1	165.0	246.7	266.8	246.2	265.5	183.5	282.7	232.7	236.2	138.7	144.5
哈尔滨	119.7	120.8	240.6	219.4	225.6	296.4	227.2	267.7	221.1	201.5	128.7	94.7
上海	113.1	144.0	139.3	182.8	178.7	152.0	258.6	222.8	134.6	188.9	147.5	102.7
南京	120.6	157.7	164.8	195.6	176.9	186.1	229.3	218.8	180.3	218.4	181.1	111.7
杭州	88.9	147.8	116.5	176.3	180.9	121.7	286.4	201.1	120.1	200.3	145.2	89.7
合肥	116.2	186.0	172.1	205.4	158.3	183.4	232.1	175.5	179.6	167.9	157.0	94.2
福州	85.4	156.4	91.7	142.9	163.2	192.3	266.0	216.4	148.9	242.5	137.6	125.5
南昌	83.5	146.5	121.7	193.6	194.5	143.6	289.6	231.4	187.9	245.8	157.0	153.9
济南	183.6	227.9	196.0	219.7	246.3	164.4	151.2	126.7	176.4	192.6	184.0	110.9
青岛	176.8	214.3	223.9	230.0	211.7	155.1	158.4	137.5	214.4	216.0	165.9	118.9
郑州	116.6	168.0	155.1	191.6	237.5	167.0	123.0	100.5	152.0	133.6	144.4	74.9
武汉	90.1	172.6	144.3	187.7	165.6	149.4	221.1	157.0	184.0	172.3	157.7	116.2
长沙	44.3	89.6	105.4	160.2	165.0	129.2	237.5	150.8	159.7	163.3	105.7	115.5
广州	121.7	104.4	55.3	83.4	125.0	176.5	155.8	167.1	185.0	146.5	165.1	181.6
南宁	48.4	79.8	53.3	123.4	136.5	197.0	133.3	191.3	187.8	192.3	120.2	153.3
海口	63.5	108.5	88.1	164.6	220.1	256.2	217.0	243.8	193.8	210.5	157.6	145.5
桂林	37.2	101.3	66.8	114.1	124.8	160.4	147.7	202.1	210.5	189.8	132.4	140.3
重庆	9.0	45.0	49.2	150.5	99.6	99.1	190.9	175.4	80.2	33.0	38.5	4.3
温江	34.0	74.0	70.0	145.0	118.0	85.0	139.0	111.0	71.0	43.0	52.0	42.0
贵阳	11.9	64.3	53.1	118.9	93.1	71.0	126.8	130.0	103.6	57.6	53.9	104.8
昆明	230.1	215.3	234.2	207.1	196.8	122.6	151.1	151.3	127.1	146.7	148.7	173.1
拉萨	245.5	249.8	271.2	214.9	288.8	252.2	172.9	224.1	258.3	256.0	252.6	254.0
西安	111.8	156.8	122.3	223.1	250.9	215.0	177.7	141.3	142.7	143.8	131.1	96.6
兰州	183.0	208.0	214.0	255.0	263.0	252.0	265.0	226.0	175.0	195.0	173.0	156.0
西宁	180.1	178.7	184.6	257.8	206.1	229.5	236.3	194.5	144.0	179.4	212.7	186.1
银川	164.3	226.5	224.5	271.1	278.9	278.9	285.9	231.8	229.9	232.1	188.0	139.7
乌鲁木齐	89.1	145.4	153.2	258.5	284.0	309.7	320.8	292.8	266.1	223.7	95.0	57.3

利用分层聚类过程将文件中的各个月份按日照时数聚类。可以看到，其中的聚类分析变量都有相同的单位，这里就不做标准化处理了。执行以下操作：

执行【Analyze】/【Classify】/【Hierarchical Cluster】命令，弹出对话框

Variables : Jan、Feb、Mar、.....、Dec

Cluster : Variables

单击【Plots】按钮

勾选 Dendrogram

单击【Continue】按钮

单击【OK】按钮

将聚类分析变量移入【Variables】框
对变量进行聚类

弹出【Plots】子对话框

输出龙骨图

回到主对话框

生成以下结果

表 11-11 是观测量概述表。

表 11-11 观测量概述表

Case Processing Summary^a

Cases					
Valid		Missing		Total	
N	Percent	N	Percent	N	Percent
34	100.0%	0	.0%	34	100.0%

a. Squared Euclidean Distance used

表 11-12 是聚类进度表，表中列出了变量合并的详细步骤。以 Stage 为 1 时，即以第一步聚类为例，将变量 4（Cluster 1）和变量 5（Cluster 2）聚类合并，其中的距离测度系数（Coefficients）是 29568.040，这个类的下一次聚类合并（Next Stage）将出现在步骤 8。其中 Stage Cluster First Appears（复聚类首次出现的步骤数）下的 Cluster 1 和 Cluster 2 为 0 时表示非类的合并，若不为 0 则表示某个类，且显示的数字就是生成该类的那个步骤数。

表 11-12 聚类过程进度表

Agglomeration Schedule

Stage	Cluster Combined		Coefficients	Stage Cluster First Appears		Next Stage
	Cluster 1	Cluster 2		Cluster 1	Cluster 2	
1	4	5	29568.040	0	0	8
2	2	11	33070.370	0	0	5
3	9	10	40678.360	0	0	6
4	6	8	59290.740	0	0	6
5	2	3	66749.025	2	0	7
6	6	9	68261.250	4	3	8
7	1	2	71685.990	0	5	9
8	4	6	104261.998	1	6	10
9	1	12	123661.408	7	0	11
10	4	7	128036.093	8	0	11
11	1	4	210200.467	9	10	0

图 11-12 所示为垂直冰柱图，图中显示了各变量依次在不同聚类数时的分类归属情况。垂直冰柱图应从下往上看，该图左边的坐标代表聚类个数。若两根冰柱中间有空隙，则代表在对应聚类个数下，这两个变量（或观测量）是属于不同类的。相反，若两根冰柱相连，则代表这两个变量（或观测量）在该聚类个数下是属于一类的。

对于系统聚类法，由于把每个变量看作单独的一类，所以本例最开始一共有 12 类。系统聚类法的第一步是根据日照时数将四月和五月合并为一类，对应到图 11-12 可以看出，垂直冰柱图在左边聚类个数坐标等于 11 的时候，四月和五月的冰柱图已经连接在了一起，代表此时把这两个变量聚为一类。而其余各月份变量此时还单独为一类，所以冰柱图之间是有空隙的。

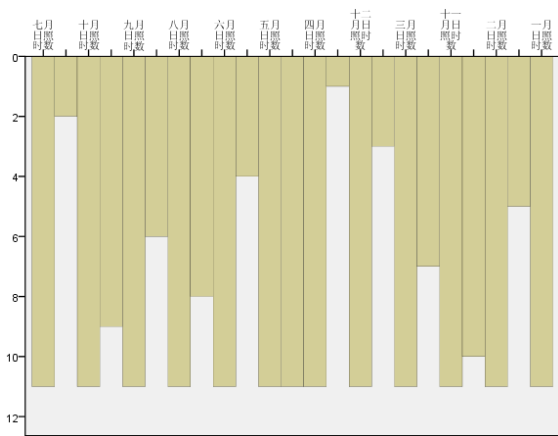


图 11-12 垂直冰柱图

图 11-13 为龙骨图，图中横向距离表示差异的大小，从图中可以清晰地看出整个变量的聚类过程。同时，如果要根据日照时数将各月份分为两类的话，那么从龙骨图可以看出 4-10 月应归为一类，其余的月份归为另外一类。

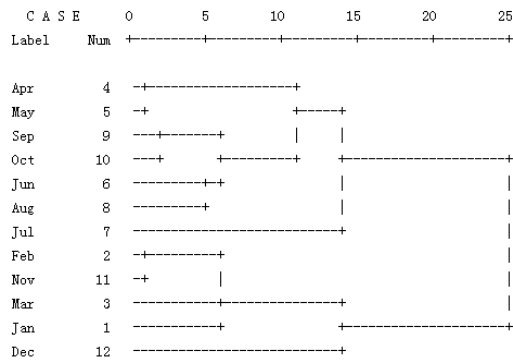


图 11-13 龙骨图

11.4 两步聚类法——TwoStep Cluster过程

两步聚类法是一种探索性的聚类方法，是随着人工智能的发展而发展起来的智能聚类方法中的一种。它主要用于解决海量数据或者具有复杂类别结构的聚类分析问题。

11.4.1 两步聚类法基本原理

两步聚类方法具有以下特点：

- 具备同时处理分类变量和连续变量的能力；
- 自动选择聚类数；
- 通过预先选取样本中的部分数据构建聚类模型，两步聚类可以处理超大样本量的数据。

两步聚类法的功能非常强大，但原理较为复杂，因此，这里只能粗略介绍它的基本原理。两步聚类法正如它的字面意思一样，是分成两个步骤完成的聚类。

STEP 01 预聚类。对记录进行初始的归类，在这一步里用户自定义最大类别数。这一步骤主要是通过构建和修改聚类特征树（CF Tree）来完成的。

STEP 02 正式聚类。对第一步完成的初步聚类进行再聚类并确定最终的聚类方案，在这一步，系统会根据一定的统计标准确定聚类的类别数目。到达这一步骤时，需要处理的类别数已经不像原始数据那样多了，可以通过传统的聚类方法进行聚类，在 SPSS 中采用的是合并型分层聚类法。

由于其有预先构建聚类模型这一特性，两步聚类会依据进入的样本随机考虑聚类数，所以它对数据进入的次序是敏感的，不同的进入次序会得到不同的聚类结果。

解决这个问题的办法包括：

- 使用一指定随机变量并以此排序，控制样本进入的顺序，再进行两步聚类，并重复几次此过程以检验聚类结果的稳定性。
- 如果样本量不是很大，又不用同时处理多类的数据，可以使用其他的经典聚类方法。

11.4.2 TwoStep Cluster过程界面操作介绍

执行【Analyze】/【Classify】/【TwoStep Cluster】命令，弹出如图 11-14 所示的【TwoStep Cluster Analysis】（两步聚类分析）对话框。

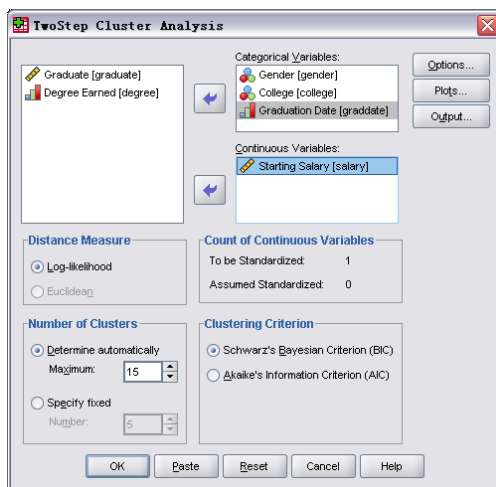


图 11-14 【两步聚类分析】对话框

1. 【Categorical Variables】框

分类变量框，用于放置分类变量，也可以放入连续变量，这时系统将把连续变量当做分类变量来处理。

2. 【Continuous Variables】框

连续变量框，用于放置连续变量，分类变量无法被移入其中。

3. 【Distance Measure】单选框

距离测量单选框，用于选择距离的测量方法。其中包含两个选项，Log-likelihood（对数似然值）和 Euclidean（欧氏距离），前者为系统默认值。当没有选入分类变量时，读者可以任意选择这两种方法中的一种，只是如果选择 Euclidean 的话，相当于使用传统聚类方法进行聚类，若有分类变量选入时，系统自动令 Euclidean 选项无法使用，就只能使用 Log-likelihood 了。

4. 【Court of Continuous Variables】框

连续变量计数框，用于显示被选入的连续变量的数目及状态。

5. 【Number of Clusters】单选框

聚类数目单选框。其中包含两个选项，第一项是指由系统自动决定分类数目，并在下面的 Maximum 栏内输入一个数值来限制分类的最大数目，此选项为系统默认选项；第二项是指由用户自己确定分类数目，并在下面的 Number 栏内输入这个指定值。

6. 【Clustering Criterion】单选框

聚类准则单选框。在两步聚类的第二步层次聚类中，系统对它的每一个阶段都会计算反映现有分类是否符合现有数据的统计指标。为此，SPSS 提供了两个准则，分别是 AIC（Akaike Information Criterion）准则和 BIC（Schwartz Bayesian Criterion）准则，这两个指标越小，聚类效果越好。系统会根据 AIC 和 BIC 的大小，以及类间最短距离的变化情况来确定最优的聚类类别数。

7. 【Options】子对话框

选项子对话框。单击【Options】按钮，弹出此子对话框，如图 11-15 所示。

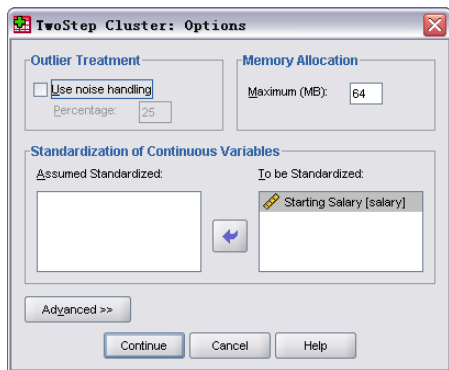


图 11-15 【Options】子对话框

① Outlier Treatment 框：设置用于建立 CF 树过程的这个算法的工具。

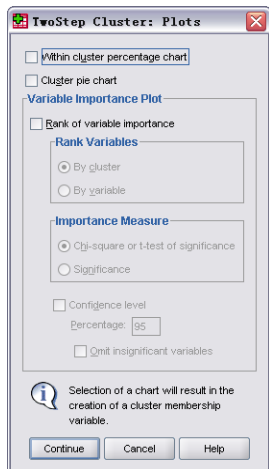
② Memory Allocation 框：内存分配框，选择算法最大的内存分配量，系统默认为 64MB，一般使用系统默认值。

③ Standardization of Continuous Variables 框：连续变量标准化框。可以将已经标准化了的连续变量从右侧的 To be Standardized 列表框中移到左侧的 Assumed Standardized 框。

④ **Advanced** 部分：高级选项部分，主要用于对前面提到的聚类特征树的选项设置。一般使用系统默认值即可。

8. 【Plots】子对话框

图形子对话框。单击【Plots】按钮，弹出此子对话框，如图 11-16 所示。其中包括 3 个部分。



• Within cluster percentage chart 选项

输出各变量在聚类中比重图。

• Cluster pie chart 选项

输出聚类饼分图。

• Variable Importance Plot 框

变量重要性图形框，用于输出一类独特的图形，用来比较各个变量对聚类结果的重要性。勾选 **Rank of variable importance** 选项，表示输出此图。

① **Rank variables** 单选框：其中有两个选项，**By cluster** 和 **By variable**，若选择 **By cluster** 选项，则为每个变量做一张条图，通过直条的长度来确定该变量对于各个类别的重要性；若选择 **By variable** 选项，则为每个类别做两张图，一张图用于比较连续变量对于聚类结果的重要性，一张图用于比较分类变量对于聚类结果的重要性，同样是通过直条的长度来确定该类别中各个变量的重要性。

② **Importance Measure** 单选框：重要性测度单选框，用于选择变量重要性的测度方法。一般使用系统默认的第一个选项。

③ **Confidence level** 选项：置信度选项，用于设置置信度。

④ **Omit insignificant variables** 选项：勾选此项，则系统将自动删除不重要的分析变量。

9. 【Output】子对话框

输出子对话框。单击【Output】按钮，弹出此子对话框，如图 11-17 所示。其中包括 3 个部分。

① **Statistics** 复选框：统计量复选框，选择要输出的统计量，系统默认为前两个选项。第一项指输出反映各类的轮廓特征的质心表和频数表；第二项指输出聚类分布表；第三项指输出自动聚类结果列表。

② **Working Data File** 框：用于在文件中创建一个新变量，保存各个观测量的所属类别。

③ **XML Files** 框：选择输出聚类的最终模型或聚类特征树到指定位置。

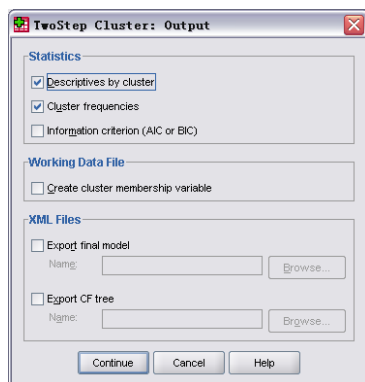


图 11-17 【Output】子对话框

11.4.3 引例及结果解释

通过一个例子来介绍【TwoStep Cluster】过程的操作及其结果。

例 11.4 以 SPSS 自带的文件“University of Florida graduate salaries.sav”为例，其中的变量包括“graduate”——毕业生代号、“gender”——性别、“college”——毕业学校、“salary”——刚毕业时的工资、“degree”——学位、“graddate”——毕业日期。

利用两步聚类法对文件中的数据进行分类。执行以下操作：

执行【Analyze】/【Classify】/【TwoStep Cluster】命令，弹出对话框	
Categorical Variables : gender、college、graddate	选入待分析的分类变量
Continuous Variables : salary	选入待分析的连续变量
单击【Plots】按钮	
勾选 Within cluster percentage chart	输出各变量在聚类中的比重图
勾选 Cluster pie chart	输出聚类饼图
勾选 Rank of variable importance	输出变量重要性图
Rank Variables : By variable	
单击【Continue】按钮	
单击【Output】按钮	
勾选 Information criterion	输出自动聚类结果列表
单击【Continue】按钮	
单击【OK】按钮	
生成以下结果	

输出结果被分成了 8 个部分，每个部分由若干表格或图形组成。

1. TwoStep Cluster部分

表 11-13 所示是自动聚类结果列表，其中列出了不同类别数的不同指标，这些指标都是用于确认最佳类别数的。表中第一列输出聚类数目；第二列输出计算的 BIC 值，即 Bayes 信息准则，这个数值是用于确认最佳类别数的重要指标，其中，数值越小代表效果越好；第三列输出反映相邻两种结果的 BIC 值之差，从表中可以看出，相比之下，聚类数为 4 类以后，BIC 值的改变就不大了；第四列输出相邻两个 BIC 差值的比例；第五列输出相邻两步的最小类间距离比，以进一步确认最佳类别数，从表中可见最小类间距离比有四个峰值，分别对应 4 类、8 类、11 类、14 类，该指数越大表示当前结果越好。根据这些指标值，系统自动认定最佳类别数，然后用该类别数进行后续的分析。

表 11-13 自动聚类结果列表

Auto-Clustering				
Number of Clusters	Schwarz's Bayesian Criterion (BIC)	BIC Change ^a	Ratio of BIC Changes ^b	Ratio of Distance Measures ^c
1	8292.675			
2	7094.380	-1198.295	1.000	1.092
3	6005.065	-1089.315	.909	1.667
4	5388.117	-616.948	.515	1.771
5	5079.286	-308.831	.258	1.067

续表

Number of Clusters	Schwarz's Bayesian Criterion (BIC)	BIC Change ^a	Ratio of BIC Changes ^b	Ratio of Distance Measures ^c
6	4795.607	-283.679	.237	1.052
7	4530.601	-265.006	.221	1.001
8	4265.957	-264.644	.221	1.118
9	4038.978	-226.979	.189	1.079
10	3835.210	-203.767	.170	1.247
11	3689.768	-145.442	.121	1.236
12	3589.419	-100.349	.084	1.150
13	3514.064	-75.354	.063	1.066
14	3448.978	-65.086	.054	1.102
15	3398.328	-50.650	.042	1.002

a. The changes are from the previous number of clusters in the table.

b. The ratios of changes are relative to the change for the two cluster solution.

c. The ratios of distance measures are based on the current number of clusters against the previous number of clusters.

表 11-14 所示是类别分布表，表中列出了每个类别所包含的观测量数目。可见，系统认定类别数为 4 类，每类中包含的观测量数目都相差不大。

表 11-14 类别分布表

Cluster Distribution				
		N	% of Combined	% of Total
Cluster	1	336	30.5%	30.5%
	2	228	20.7%	20.7%
	3	300	27.3%	27.3%
	4	236	21.5%	21.5%
	Combined	1100	100.0%	100.0%
Total		1100		100.0%

2. Cluster Profiles部分

输出各类的轮廓特征。

表 11-15 所示为质心表，用于输出连续变量在每个类别中的均值和标准差。

表 11-15 质心表

Centroids			
		Starting Salary	
		Mean	Std. Deviation
Cluster	1	25086.31	7714.591
	2	24839.47	7905.729
	3	24379.42	4448.036
	4	30781.36	5280.409
	Combined	26064.20	6967.982

3. Frequencies部分

表 11-16 和表 11-17 都是分类变量频数表。这里省略了变量“college”的频数表。表中列出了不同变量的分组在各类中的频数分布。以表 11-16 为例，可以看出，女性主要被分配到了第一类和第三类中。其中，第一类最多，71.6%的女性都被划分到第一类。而男性则分别在第二、三、四类。

表 11-16 分类变量 gender 的频数表

		Gender			
		Female		Male	
		Frequency	Percent	Frequency	Percent
Cluster	1	336	71.6%	0	.0%
	2	0	.0%	228	36.1%
	3	133	28.4%	167	26.5%
	4	0	.0%	236	37.4%
	Combined	469	100.0%	631	100.0%

表 11-17 分类变量 graddate 的频数表

		Graduation Date							
		Fall 89		Spring 90		Fall 90		Spring 91	
		Frequency	Percent	Frequency	Percent	Frequency	Percent	Frequency	Percent
Cluster	1	68	26.4%	124	31.4%	61	27.2%	83	37.2%
	2	56	21.7%	73	18.5%	35	15.6%	64	28.7%
	3	78	30.2%	104	26.3%	87	38.8%	31	13.9%
	4	56	21.7%	94	23.8%	41	18.3%	45	20.2%
	Combined	258	100.0%	395	100.0%	224	100.0%	223	100.0%

4. Attribute Importance部分

图 11-18 所示为聚类饼图，将每一类别用饼图的形式表示，实际上就是聚类分布表（表 11-14）的图形表示。

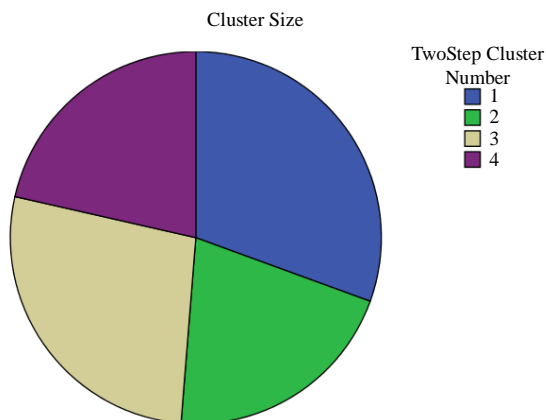


图 11-18 聚类饼分图

5. Within Cluster Percentage部分

图 11-19~图 11-21 都是分类变量在各个类别中的占比图。以图 11-19 为例，可以看出，在第一类中全为女性，占比为 100%，第二、第四类全为男性，第三类中男女比例基本相当，女性数略多。

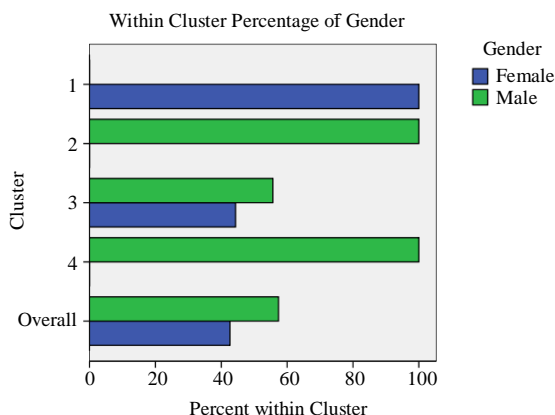


图 11-19 变量 gender 在聚类中的比重图

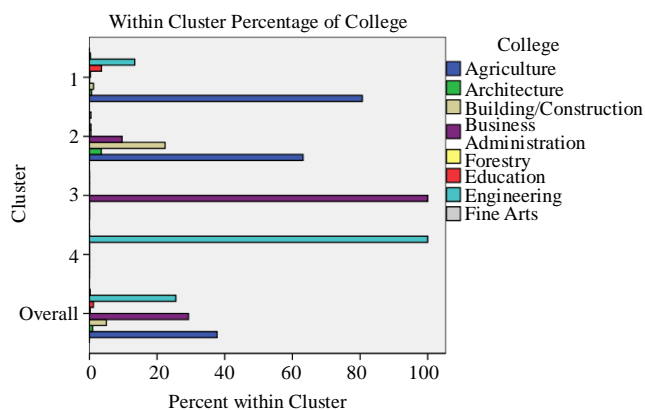


图 11-20 变量 college 在聚类中的比重图

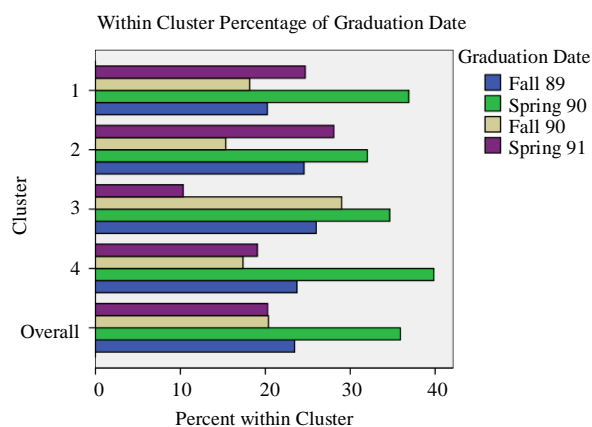


图 11-21 变量 graddate 在聚类中的比重图

6. Within Cluster Variation部分

图 11-22 所示为连续变量在各个类别中的误差图，它是用图形来表示均值和 95%置信区间的范围。它也就是质心表（表 11-15）的图形表示。

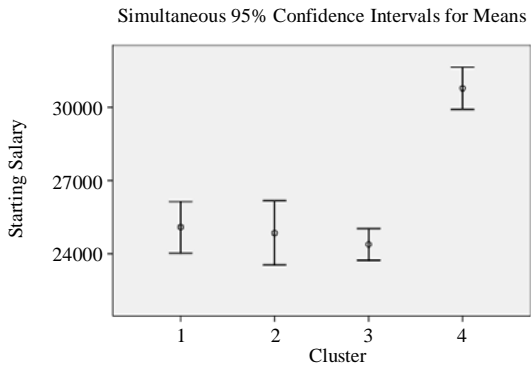


图 11-22 误差图

7. Categorical Variablewise Importance部分

图 11-23~图 11-26 分别为 4 个类别中的分类变量重要性图。图中用直条的长度和方向来表示各个变量在每一类中的重要性。

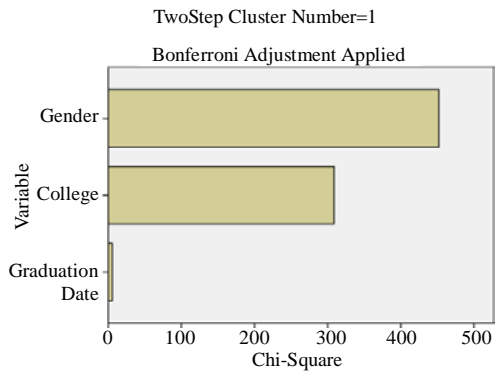


图 11-23 类别 1 中分类变量重要性图

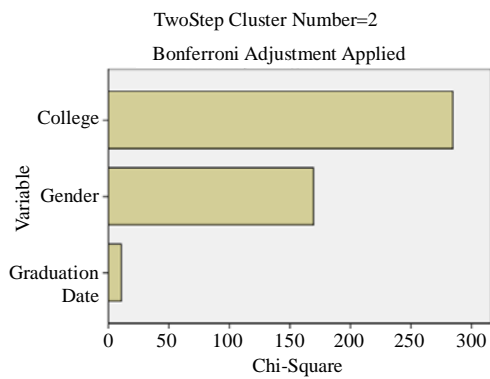


图 11-24 类别 2 中分类变量重要性图

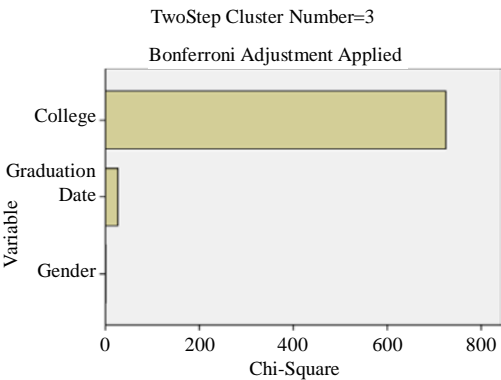


图 11-25 类别 3 中分类变量重要性图

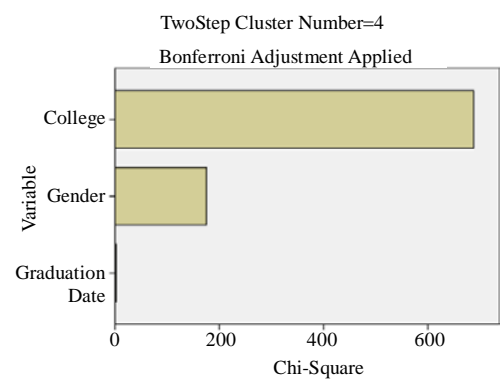


图 11-26 类别 4 中分类变量重要性图

8. Continuous Variablewise Importance部分

与 Categorical Variablewise Importance 部分图形类似，但由于本例中只有一个连续变量，因此输出无意义。

11.5 判别分析——Discriminant过程

判别分析在实际应用中非常具有实际意义，下面介绍判别分析的【Discriminant】过程。

11.5.1 判别分析基本原理

根据判别准则的不同，不同判别分析的算法也有很大的差异，但它们都有一个一般过程。

STEP 01 计算需要用到的一些反映样品特征的值，比如均值、协方差阵，等等。

STEP 02 根据一定的原则建立判别函数。 $y = c_1x_1 + c_2x_2 + \cdots + c_nx_n$ 为判别函数的一般形式，其中 y 是判别指标， x_1, x_2, \cdots, x_n 为反映研究对象特征的变量， c_1, c_2, \cdots, c_n 为判别系数。建立判别函数就是要确定这些系数。

STEP 03 确定判别准则。有的判别准则需要计算一些判别时用到的参数，比如 Fisher 判别需要计算临界值。

STEP 04 检验判别效果，即验证判别函数用来进行判别时的准确度。

STEP 05 对待判样品判别归类。

下面介绍判别分析中最常用到的四种判别法的基本思想和特点。

(1) 距离判别法：根据已知分类的数据，分别计算各类的均值（重心），判别准则是任意给一次观测，若它与第 i 类的重心距离最近，就认为它来自第 i 类。这里的距离一般采用马氏距离。距离判别适合对自变量均为连续变量的情况进行分类，它对各类的分布无特定的要求。

(2) Fisher 判别法：借助方差分析的思想构造一个判别函数，其中判别系数的确定原则是使得类间的区别最大，而且类内的离差最小，利用判别函数计算出待判样品的判别指标，然后与判别临界值进行比较，判别它的类属。Fisher 判别应用范围较广，而且对各类分布、方差都没有什么限制。但当总体个数较多时，计算比较麻烦。

(3) Bayes 判别法：在考虑先验概率的前提下，利用 Bayes 公式计算样品来自第 i 类的后验概率，使用错判损失最小的概念作判别准则，建立判别函数，将待判样品归入来自概率最大的类。Bayes 判别主要用于多类判别，它要求总体呈多元正态分布。

(4) 逐步判别法：逐步判别法与逐步回归法的基本思想类似，都是逐步引入变量，每引入一个“最重要”的变量进入判别式，同时也考虑较早引入判别式的某些变量，若其判别能力不显著了，应及时从判别式中剔除，直到判别式中没有不重要的变量需要剔除，并且也没有重要的变量要引入为止。这个筛选过程实质就是做假设检验。

在上面的介绍中，都没有涉及具体的算法过程、重要指标的数学表达式等细节，读者可以参看多元统计分析的相关教材。

11.5.2 Discriminant过程界面操作介绍

执行【Analyze】/【Classify】/【Discriminant】命令，弹出如图 11-27 所示的【Discriminant Analysis】（判别分析）对话框。

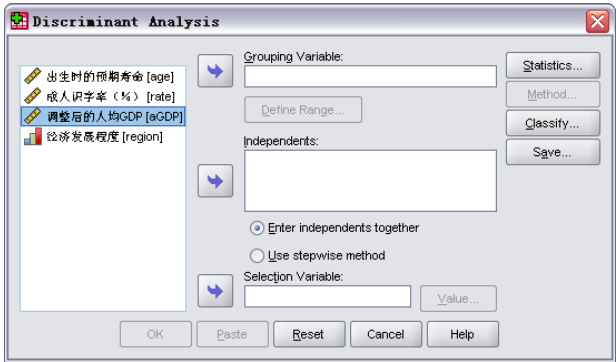


图 11-27 【判别分析】对话框

1. 【Grouping Variable】框

分组变量框，放入分组变量。选入变量后，单击【Define Range】按钮，弹出如图 11-28 所示的子对话框，用于设置变量的取值范围。

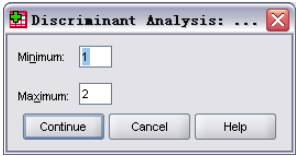


图 11-28 【Define Range】子对话框

2. 【Independents】框

自变量框，用于选入建立判别函数所需的变量。可以放入可疑变量，利用逐步法进行筛选。

3. 选项“Enter independents together”和“Use stepwise method”

前者意思为所有自变量同时进入判别函数，这是系统默认值；后者为使用逐步判别法，其实质和多元回归分析的逐步法等价。

4. 【Selection Variable】框

选择变量框，用于定义记录选择条件。选入变量以后，单击【Value】按钮，弹出一个【Set Value】子对话框，在对话框内输入一个数，表示全部记录中只有该变量取值等于这个数的记录才用于分析。

5. 【Statistics】子对话框

统计量子对话框。单击【Statistics】按钮，弹出此子对话框，如图 11-29 所示。其中有 3 个复选框。

- Descriptives 复选框

描述统计量复选框。选项 Means 的意思是输出各自变量在各类中的观测量和全部观测量的均值、标准差；选项 Univariate ANOVAs 是针对所有自变量进行单因素方差分析；选

项 Box's M 意思是输出对各类协方差矩阵相等的假设进行 Box's M 检验的结果。

- **Function Coefficients** 复选框

判别函数系数复选框。在默认情况下系统给出的是采用 Bayes 方法建立的判别函数的标准化系数。选项 Fisher's 意思是给出 Bayes 判别准则的判别函数；选项 Unstandardized 意思是给出 Fisher 判别法建立的判别函数的未标准化系数。

- **Matrices** 复选框

矩阵复选框。其中包括类内相关矩阵、类内协方差矩阵、对每一类分别显示协方差矩阵，以及总样本的协方差矩阵。

6. 【Stepwise Method】子对话框

逐步判别法子对话框。单击【Method】按钮，弹出此子对话框，如图 11-30 所示。只有在主对话框中选择了“Use stepwise method”时，此项才被激活。其中有两个单选框和一个复选框。

- **Method** 单选框

方法单选框，用于选择逐步判别分析时所用的拟合方法。系统默认选项 Wilks' Lambda 法，即每步选择 Wilk 的 λ 统计量值最小的变量进入判别函数；选项 Unexplained variance 是指每步选择类间不可解释的方差和最小的变量进入判别函数；选项 Mahalanobis' distance 是指每步选择邻近类间 Mahalanobis 距离最大的变量进入判别函数；选项 Smallest F ratio 是指每步选择根据类间 Mahalanobis 距离计算的“最小 F 比”达到最大的变量进入判别函数；选项 Rao's V 是指每步选择使 Rao's V 值的增量最大化的变量进入判别函数。

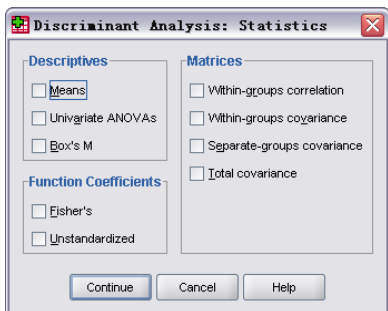


图 11-29 【Statistics】子对话框

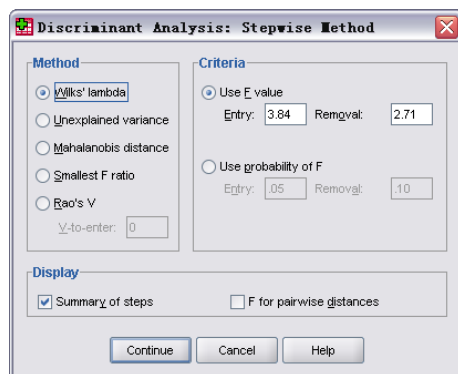


图 11-30 【Stepwise Method】子对话框

- **Criteria** 单选框

临界值单选框，用于决定终止逐步判别的临界值。可以使用 F 值或者 P 值作为标准，并在 Entry 栏和 Removal 栏内分别输入引入变量和剔除变量的临界值，使用 F 值时，要求引入变量的临界值要大于剔除变量的临界值，使用 P 值时则相反，要求引入变量的临界值要小于剔除变量的临界值。

- **Display** 复选框

显示复选框，用于选择每一步需要输出的统计量。其中包括有该步骤的汇总表，以及每一对类之间的 F 比值矩阵。

7. 【Classification】子对话框

分类子对话框。单击【Classify】按钮，弹出此子对话框，如图 11-31 所示。其中有两个单选框、两个复选框和一个选项 Replace missing values with mean，意思是在分类阶段用自变量的均值代替缺失值。

- Prior Probabilities 单选框

先验概率单选框，用于设定判别函数的先验概率。系统默认为第一个选项 All groups equal，即各类先验概率均相等；第二项是指基于各类样本量占总样本量的比例计算先验概率。

- Use Covariance Matrix 单选框

使用协方差矩阵单选框。系统默认为第一项，使用合并类内协方差矩阵进行分类；第二项是指使用各类协方差矩阵进行分类。

- Display 复选框

显示复选框，用于选择一些可以输出的指标。第一项指输出每个观测测量判别后的所属类别，并可以在下面的 Limit cases to first 栏后输入一个数，来限制输出分类结果的观测测量数目；第二项指输出分类小结表，对每一类输出判定正确和错判的观测测量数；第三项指对于每一个观测测量，输出依据除它之外的其他观测测量导出的判别函数的分类结果。

- Plots 复选框

图形复选框。选项 Combined-groups 是指生成全部类的散点图，若只有一个判别函数，则生成直方图；选项 Separate-groups 是指对每一类分别生成散点图，若只有一个判别函数，则生成直方图；选项 Territorial map 是指生成根据判别函数值将观测测量分到各类去的边界图，若只有一个判别函数，则不显示此图。

8. 【Save】子对话框

保存子对话框。单击【Save】按钮，弹出此子对话框，如图 11-32 所示。此对话框用于选择建立新变量将判别分析结果保存到当前工作文件中。其中包括三个选项，分别表示建立新变量保存预测观测测量所属类的值、保存判别指数，以及保存各观测测量属于各类的概率值。

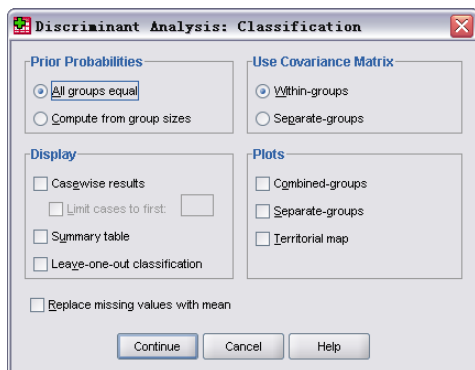


图 11-31 【Classification】子对话框

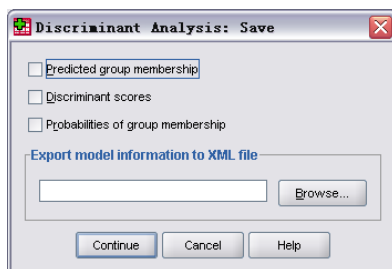


图 11-32 【Save】子对话框

11.5.3 引例及结果解释

通过一个例子介绍【Discriminant】过程的操作及其结果。

例 11.5 以数据文件“1995 年人类发展报告部分数据.sav”为例，表 11-18 列出了此文件内容，其中选取了高发展水平的国家和中等发展水平的国家各 5 个作为两组样品，另选了 4 个国家作为待判别样品。（数据来源：《多元统计分析》，中国统计出版社）。

表 11-18 1995 年人类发展报告部分数据

类 别 region	国家名称 name	出生时的预期寿命（岁） age	成人识字率（%） rate	调整后的人均 GDP aGDP
第一类 （高发展 水平 国家）	美国	76	99	5374
	日本	79.5	99	5359
	瑞士	78	99	5372
	阿根廷	72.1	95.9	5242
	阿联酋	73.8	77.7	5370
第二类 （中等发展 水平 国家）	保加利亚	71.2	93	4250
	古巴	75.3	94.9	3412
	巴拉圭	70	91.2	3390
	格鲁吉亚	72.8	99	2300
	南非	62.9	80.6	3799
待判 样品	中国	68.5	79.3	1950
	罗马尼亚	69.9	96.9	2540
	希腊	77.6	93.8	5233
	哥伦比亚	69.3	90.3	5158

利用逐步判别法判别待判样品的类别。其中待判样品的“region”变量值为缺失值。执行以下操作：

执行【Analyze】/【Classify】/【Discriminant】命令，弹出【Discriminant Analysis】对话框	
Grouping Variable：region	选入分组变量
单击【Define Range】按钮	弹出【Define Range】子对话框
Minimum：1	设置分组变量取值范围
Maximum：2	
单击【Continue】按钮	回到主对话框
Independents：age、rate、aGDP	选入自变量
选择 Use stepwise method	使用逐步判别法
单击【Save】按钮	弹出【Save】子对话框
勾选 Predicted group membership	建立新变量保存观测值预测分组
单击【Continue】按钮	回到主对话框
单击【OK】按钮	生成以下结果

输出结果被分成了 5 个部分，每个部分由若干表格或图形组成。

1. Discriminant部分

其中包括两个报表，分别为记录纳入情况简报和各组例数报告，这里就不再赘述。

2. Stepwise Statistics部分

由 4 个表格组成，这 4 个表格都是逐步判别分析的运行记录，可以参看多元回归分析章节。从表 11-19 中可以看出第一步就纳入变量“aGDP”，在表 11-22 给出的 Wilk's Lambda 检验结果中，可以看到 Sig.值<0.05，即检验结果是拒绝 H_0 ，这说明变量“aGDP”对正确判断分析是有作用的。

表 11-19 被引入/剔除的变量表

Variables Entered/Removed ^{a,b,c,d}									
Step	Entered	Wilks' Lambda							
		Statistic	df1	df2	df3	Exact F			
						Statistic	df1	df2	Sig.
1	调整后的人均 GDP	.225	1	1	8.000	27.482	1	8.000	.001

At each step, the variable that minimizes the overall Wilks' Lambda is entered.

- Maximum number of steps is 6.
- Minimum partial F to enter is 3.84.
- Maximum partial F to remove is 2.71.
- F level, tolerance, or VIN insufficient for further computation.

表 11-20 参加判别分析的变量表

Variables in the Analysis		
Step	Tolerance	F to Remove
1 调整后的人均 GDP	1.000	27.482

表 11-21 未参加判别分析的变量表

Variables Not in the Analysis				
Step	Tolerance	Min. Tolerance	F to Enter	Wilks' Lambda
0 出生时的预期寿命	1.000	1.000	4.787	.626
成人识字率(%)	1.000	1.000	.213	.974
调整后的人均 GDP	1.000	1.000	27.482	.225
1 出生时的预期寿命	.923	.923	2.835	.160
成人识字率(%)	.887	.887	1.103	.195

表 11-22 Wilk's Lambda 检验结果表

Wilks' Lambda									
Step	Number of Variables	Lambda	df1	df2	df3	Exact F			
						Statistic	df1	df2	Sig.
1	1	.225	1	1	8	27.482	1	8.000	.001

3. Summary of Canonical Discriminant Functions部分

由 5 个表格组成。从表 11-23 可以看出，本例仅有一个判别函数用于分析，特征值为 3.435，方差百分比为 100%，方差累计百分比为 100%，正则相关系数为 0.880。

表 11-23 特征值表

Eigenvalues				
Function	Eigenvalue	% of Variance	Cumulative %	Canonical Correlation
1	3.435 ^a	100.0	100.0	.880

a. First 1 canonical discriminant functions were used in the analysis.

表 11-24 所示是对判别函数的显著性检验，其中 Wilk's Lambda 的值等于 0.225，卡方统计量值为 11.172，自由度为 1，显著性概率为 0.001，从而认为判别函数有效。

表 11-24 Wilk 的 λ 值表

Wilks' Lambda				
Test of Function(s)	Wilks' Lambda	Chi-Square	df	Sig.
1	.225	11.172	1	.001

由表 11-25 可知，判别函数为 $F1 = 1.000 \text{ aGDP}$ ，根据这个判别函数代入各变量的数值可以计算出判别指数。

表 11-25 判别函数的标准化系数表

Standardized Canonical Discriminant Function Coefficients	
	Function
	1
调整后的人均 GDP	1.000

表 11-26 的结构矩阵是按照绝对值大小排列的各变量与主成分间的相关系数，表明判别变量与判别函数之间的相关性。

表 11-26 结构矩阵表

Structure Matrix	
	Function
	1
调整后的人均 GDP	1.000
成人识字率 (%) a	-.337
出生时的预期寿命 a	-.277

Pooled within-groups correlations between discriminating variables and standardized canonical discriminant functions
Variables ordered by absolute size of correlation Within function.

a. This variable not used in the analysis.

表 11-27 给出的是各类重心在空间中的坐标位置，即利用判别函数在各类均值处的判别指数值。

表 11-27 判别函数类型表

Functions at Group Centroids	
经济发展程度	Function
	1
高发展水平国家	1.658
中等发展水平国家	-1.658

Unstandardized canonical discriminant
functions evaluated at group means

4. Classification Statistics部分

由两个表格组成。表 11-28 给出的是缺失值报告，从表中可以看出所有的观测量都已分好类了。

表 11-28 分类结果概述表

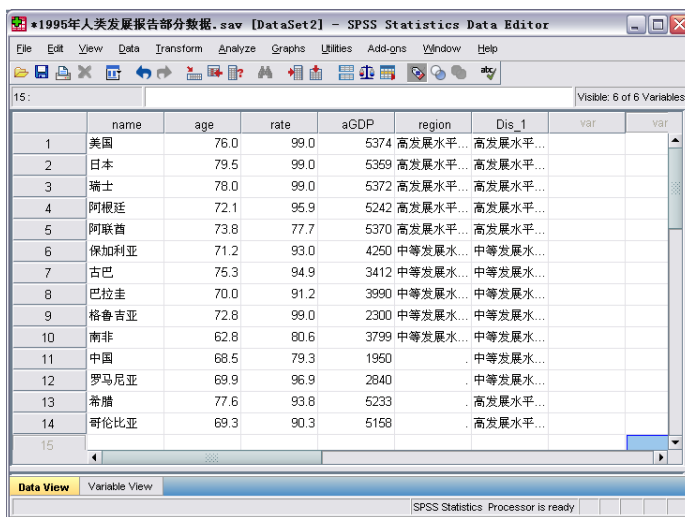
Classification Processing Summary		
Processed		14
Excluded	Missing or out-of-range group codes	0
	At least one missing discriminating variable	0
Used in output		14

表 11-29 给出各类的先验概率，按照系统默认方法，每一类的先验概率都为 0.5。

表 11-29 各类先验概率表

Prior Probabilities for Group			
经济发展程度	Prior	Cases Used in Analysis	
		Unweighted	Weighted
高发展水平国家	.500	5	5.000
中等发展水平国家	.500	5	5.000
Total	1.000	10	10.000

执行以上操作之后，原始数据文件变成如图 11-33 所示。从图中可以看出，新生成了一列保存各国的判别分析结果。可见，根据判别分析，将中国和罗马尼亚归为中等发展水平国家，而将希腊和哥伦比亚归为高发展水平国家。



	name	age	rate	aGDP	region	Dis_1	var	var
1	美国	76.0	99.0	5374	高发展水平...	高发展水平...		
2	日本	79.5	99.0	5359	高发展水平...	高发展水平...		
3	瑞士	78.0	99.0	5372	高发展水平...	高发展水平...		
4	阿根廷	72.1	95.9	5242	高发展水平...	高发展水平...		
5	阿联酋	73.8	77.7	5370	高发展水平...	高发展水平...		
6	保加利亚	71.2	93.0	4250	中等发展水...	中等发展水...		
7	古巴	75.3	94.9	3412	中等发展水...	中等发展水...		
8	巴拉圭	70.0	91.2	3990	中等发展水...	中等发展水...		
9	格鲁吉亚	72.8	99.0	2300	中等发展水...	中等发展水...		
10	南非	62.8	80.6	3799	中等发展水...	中等发展水...		
11	中国	68.5	79.3	1950	中等发展水...	中等发展水...		
12	罗马尼亚	69.9	96.9	2840	中等发展水...	中等发展水...		
13	希腊	77.6	93.8	5233	高发展水平...	高发展水平...		
14	哥伦比亚	69.3	90.3	5158	高发展水平...	高发展水平...		
15								

图 11-33 判别分析结果

11.6 本章小结

本章介绍了聚类分析与判别分析在 SPSS 中的实现，详细介绍了以下几个过程：

- K-means Cluster 过程，K-均值聚类分析；
- Hierarchical Cluster 过程，系统聚类法；
- TwoStep Cluster 过程，两步聚类法；
- Discriminant 过程，判别分析。

其中，对于三种聚类分析的方法，其主要特点归纳总结如表 11-30 所示。

表 11-30 三种聚类方法总结

聚类方法	SPSS 中的实现	运算速度	聚类对象	聚类对象性质
快速聚类	K-means Cluster 过程	快	观测量	连续变量
系统聚类	Hierarchical Cluster 过程	一般	观测量/变量	分类变量/连续变量
两步聚类	TwoStep Cluster 过程	慢	观测量/变量	分类变量/连续变量

用户在具体的实现过程中，应当根据不同的需求，结合各种方法的特点，选取适合的方法。同时，一定要结合实际背景，来检验当前聚类分析和判别分析的结果是否有实际意义。

第 12 章 因子分析与对应分析

因子分析和对应分析都是多元统计分析中的重要统计方法。在回归分析中，可以看到变量的共线性有时候会对分析产生极大地影响，因子分析和对应分析就是用来解决这类问题的。因子分析和对应分析都是将具有错综复杂关系的变量或者样品综合为数量较少的几个因子，以再现原始变量与因子之间的相互关系，同时还可以对变量或观测量进行分类。

SPSS 的【Dimension Reduction】子菜单提供了 3 个关于因子分析和对应分析的子菜单。本章将通过实际例子来学习因子分析和对应分析的一般步骤及其在 SPSS 中的实现过程。本章内容包括：

- 因子分析——Factor Analysis 过程
- 简单对应分析——Correspondence Analysis 过程
- 最优尺度分析——Optimal Scaling 过程初步认识

12.1 因子分析——Factor Analysis过程

因子分析的形成和发展已经有相当长的历史了，最早用于研究解决心理学和教育学方面的问题，目前这一方法的应用范围已十分广泛，在经济学、社会学、考古学、生物学、医学、地质学，以及体育科学等各个领域都取得了显著的成绩。

12.1.1 因子分析基本原理

介绍因子分析之前，先来说说主成分分析，因为两者非常近似。主成分分析是将多个指标化为少数相互无关的综合指标的统计方法。通常数学上的处理就是将原来 p 个指标做线性组合，作为新的综合指标，记第一个综合指标为 F_1 ，选取这个线性组合的原则是令 $Var(F_1)$ ，即 F_1 的方差最大，这时，称 F_1 为第一主成分。然后选取第二主成分 F_2 ，同样的选取原则，只是还要加上一个条件 $Cov(F_1, F_2) = 0$ ，依此类推，构造出剩下的主成分。

因子分析是主成分分析的推广和发展。它的基本思想是通过变量（或样品）的相关系数矩阵（对样品是相似系数矩阵）内部结构的研究，找出能控制所有变量（或样品）的少数几个随机变量去描述多个变量（或样品）之间的相关（相似）关系。但在这里，这少数几个随机变量是不可观测的，通常称为因子。然后根据相关性（或相似性）的大小把变量（或样品）分组，使得同组内的变量（或样品）之间相关性（或相似性）较高，但不同组的变量相关性（或相似性）较低。

按照分析的内容分类, 因子分析被分为 R 型因子分析 (对变量作因子分析) 和 Q 型因子分析 (对样品作因子分析)。两者的计算过程都是一样的, 只是出发点不同, 这里就以 R 型因子分析为主进行介绍。

R 型因子分析的数学模型

$$\begin{cases} X_1 = a_{11}F_1 + a_{12}F_2 + \cdots + a_{1m}F_m + \varepsilon_1 \\ X_2 = a_{21}F_1 + a_{22}F_2 + \cdots + a_{2m}F_m + \varepsilon_2 \\ \vdots \\ X_p = a_{p1}F_1 + a_{p2}F_2 + \cdots + a_{pm}F_m + \varepsilon_p \end{cases}$$

简化为矩阵形式就是 $X = A F + \varepsilon$, 且满足:

- ① m, p ;
- ② $Cov(F, \varepsilon) = 0$ 即 F 和 ε 是不相关的;
- ③ $D(F) = I_m$ (单位阵), 即 $F_1 \cdots F_m$ 不相关且方差皆为 1;

$$D(\varepsilon) = \begin{pmatrix} \sigma_1^2 & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & \sigma_p^2 \end{pmatrix}, \text{ 即 } \varepsilon_1 \cdots \varepsilon_p \text{ 不相关且方差不同。}$$

其中 X 是由可实测的 p 个指标所构成的 p 维随机向量, F 是不可观测的向量, 被称为公共因子或潜因子, 即前面所说的综合变量, 可以把它们理解为在高维空间中的互相垂直的 m 个坐标轴; 称 a_{ij} 为因子载荷是第 i 个变量在第 j 个公共因子上的负荷, 如果把变量 X_i 看成 m 维因子空间中的一个向量, 则 a_{ij} 表示 X_i 在坐标轴 F_j 上的投影, 矩阵 A 称为因子载荷矩阵; 称 ε 为 X 的特殊因子。

下面介绍因子模型中几个重要概念的统计意义。

1. 因子载荷

实际上 $a_{ij} = r_{X_i F_j}$, 即第 i 个变量与第 j 个公共因子的相关系数, 它的统计意义就是第 i 个变量在第 j 个公共因子上的负荷, 反映了第 i 个变量在第 j 个公共因子上的相对重要性。

2. 变量共同度

所谓变量 X_i 的共同度定义为因子载荷矩阵 A 中第 i 行元素的平方和, 即

$$h_i^2 = \sum_{j=1}^m a_{ij}^2 \quad i = 1, \cdots, p$$

共同度刻画全部公共因子对变量 X_i 的总方差所做的贡献, 它越接近 1, 说明该变量的几乎全部原始信息都被所选取的公共因子说明了; 若它接近 0, 说明公共因子对 X_i 的影响很小, 主要由特殊因子来描述。

3. 公共因子 F_j 的方差贡献

它的定义为因子载荷矩阵中各列元素的平方和, 记为

$$S_j = \sum_{i=1}^p a_{ij}^2 \quad j = 1, \cdots, m$$

它表示同一公共因子 F_j 对所有变量所提供的方差贡献的总和，它是衡量公共因子相对重要性指标。

4. 因子旋转

因子旋转的目的是为了使得因子载荷矩阵的结构简化，便于对公共因子进行解释，这里所谓的结构简化是使每个变量仅在一个公共因子上有较大的载荷，而在其余公共因子上载荷比较小。这种变换因子载荷矩阵的方法称为因子轴的旋转。旋转的方法有很多种，如正交旋转，斜交旋转等。

5. 因子得分

因子分析的数学模型是将变量表示成公共因子的线性组合，而在实际应用中，往往需要用公共因子代表原始变量，即将公共因子表示为变量的线性组合，即

$$F_j = \beta_{j1}X_1 + \cdots + \beta_{jp}X_p \quad j=1, \cdots, m \quad (12.1)$$

式 (12.1) 就是因子得分函数，用它来计算每个样品的公共因子得分。估计因子得分有很多种方法，比如加权最小二乘法，回归法等。

下面介绍因子分析的一般步骤。

STEP 01 将原始数据标准化。

STEP 02 建立变量的相关系数矩阵 R 。

STEP 03 求 R 的特征根及相应的单位特征向量，根据累计贡献率要求，取前 m 个特征根及相应的特征向量，写出因子载荷矩阵 A 。

STEP 04 对 A 施行因子旋转。

STEP 05 计算因子得分。

12.1.2 Factor Analysis过程界面操作介绍

执行【Analyze】/【Dimension Reduction】/【Factor】命令，弹出如图 12-1 所示的【Factor Analysis】（因子分析）对话框。

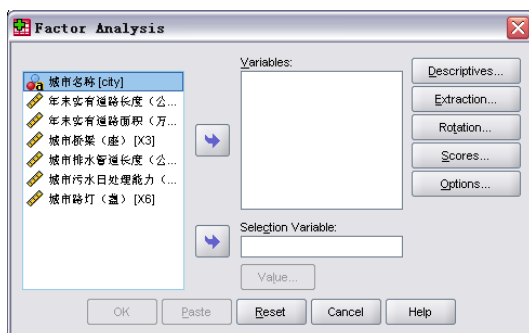


图 12-1 【因子分析】对话框

1. 【Variables】框

变量框，用于放置将用于做因子分析的变量。

2. 【Selection Variable】框

选择变量框，用于定义记录选择条件。选入变量以后，单击【Value】按钮，弹出一个【Set Value】子对话框，在对话框内输入一个数，表示全部记录中只有该变量取值等于这个数的记录才用于分析。

3. 【Descriptives】子对话框

描述子对话框，从中选择需要输出的统计量。单击【Descriptives】按钮，弹出此子对话框，如图 12-2 所示。其中有两个复选框。

• Statistics 复选框

统计量复选框。第一个选项的意思是单变量描述统计量，用于输出各个分析变量的均值、标准差，以及观测量数。第二个选项的意思是输出原始分析结果，包括原变量的公因子方差，与变量相同个数的因子，各因子的特征根及其所占总方差的百分比和累计百分比。系统默认为第二项。

• Correlation Matrix 复选框

相关矩阵复选框。其中给出的选项都是与变量间的相关性指标及相关检验有关的。

① Coefficient: 所有变量间的相关系数矩阵。

② Significance levels: 显著性水平，输出所有变量相关系数单测检验的 P 值。

③ Determinant: 相关系数矩阵的行列式值。

④ KMO and Bartlett's test of sphericity: KMO 检验和 Bartlett 球形检验。这个选项的意义将在结果分析中介绍。

⑤ Inverse: 相关系数矩阵的逆矩阵。

⑥ Reproduced: 再生相关系数矩阵。输出因子分析后的估计相关系数矩阵及残差阵。

⑦ Anti-image: 反映像协方差阵和相关阵。

4. 【Extraction】子对话框

提取子对话框，用于选择因子提取的方法。单击【Extraction】按钮，弹出此子对话框，如图 12-3 所示。其中包含以下 5 个部分。

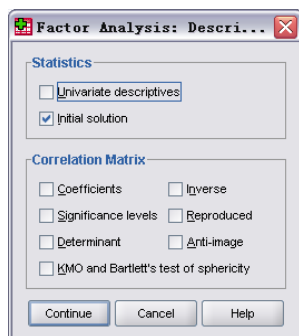


图 12-2 【Descriptives】子对话框

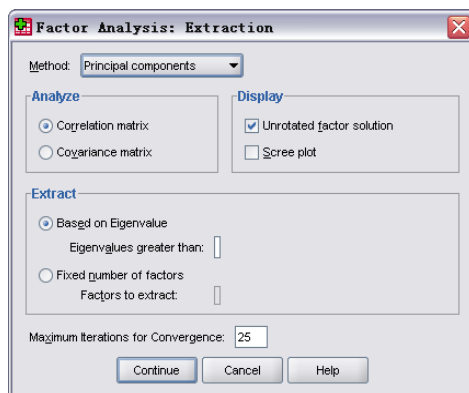


图 12-3 【Extraction】子对话框

- Method 下拉列表

用于选择公共因子的提取方法。系统提供了 7 种公共因子提取方法，系统默认为 Principal component（主成分分析法）。剩下的 6 种方法分别是不加权最小二乘法、广义最小二乘法、极大似然法、主轴因子法、 α 因子法和映像因子法。

- Analyze 单选框

用于选择使用变量间的相关矩阵还是协方差矩阵进行分析。系统默认为相关矩阵。

- Display 复选框

用于选择显示的内容。第一项 Unrotated factor solution 是指显示未经旋转变换的因子提取结果；第二项 Scree plot 是指碎石图，用于显示各因子的重要程度，这个图将在结果分析中介绍。系统默认选项为第一项。

- Extract 单选框

用于设定公共因子的提取标准。第一项意思是以特征根大于指定数值为提取标准，在后面的栏内输入指定数值，系统默认为 1；第二项意思是自定义提取因子的数量，在后面栏内输入此数量值。

- Maximum Iterations for Convergence 栏

收敛时的最大迭代次数，系统默认为 25 次。

5. 【Rotation】子对话框

旋转子对话框，用于选择因子旋转的方法。单击【Rotation】按钮，弹出此子对话框，如图 12-4 所示。其中有一个复选框和一个单选框。

- Method 单选框

提供因子旋转的 5 种方法。其中有 6 个选项，系统默认为第一个选项 None，即不进行旋转。剩下的因子旋转方法意义如下。

① Varimax：方差最大化正交旋转。这种旋转方法使每个因子仍然保持直角正交，且具有高载荷。这个方法是最常用的因子旋转方法。

② Direct Oblimin：斜交旋转法。在其下的 Delta 栏内输入参数值，规定输入的数值范围为-1 到 0.8 之间，系统默认值为 0，表示因子分析的解最倾斜。

③ Quartimax：四分旋转法。这种方法意在利用最少的因子解释每个变量。

④ Equamax：平均正交旋转法。它将 Varimax 和 Quartimax 结合起来，使得高载荷因子的变量数和需解释变量的因子数都达到最小的旋转法。

⑤ Promax：斜交旋转法。这种方法是斜交旋转中最常用的一种，它的计算速度很快，旋转后允许因子间存在相关，适合有大量数据的情况。在其下的 Kappa 栏内输入控制斜交旋转的参数，系统默认参数值为 4。

- Display 复选框

设置旋转解的输出。其中包含两个选项，第一项 Rotated solution 是指输出主成分转换矩阵，该矩阵提供旋转前后因子之间的变换系数。只有当 Method 单选框内选择了一个旋转方法后，此项才被激活；第二项 Loading plots 是指输出二维或者三维的因子载荷图。该

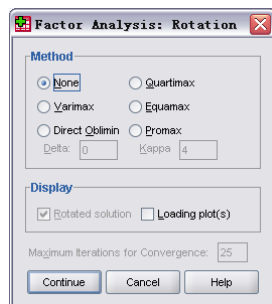


图 12-4 【Rotation】子对话框

图中坐标轴为因子值，各变量以散点的形式分布其中。若仅提取出一个公共因子，则不输出因子载荷图。

6. 【Factor Scores】子对话框

因子得分子对话框。单击【Scores】按钮，弹出此子对话框，如图 12-5 所示。其中有两个部分。

- **Save as variables 选项**

指在数据文件中建立一个新变量，用于保存各观测量的因子得分。

- **Method 框**

选择因子得分的计算方法。只有当选择 **Save as variables** 选项后，此项才被激活。其中选项包括 **Regression**（回归法）、**Bartlett**（巴特列特法）、**Anderson-Rubin**（安德森-鲁宾法），系统默认为回归法。

- **Display factor score coefficient matrix 选项**

若选择此项，则在结果中输出因子得分系数矩阵及因子得分的协方差矩阵。

7. 【Options】子对话框

在选项子对话框中，单击【Options】按钮，弹出此子对话框，如图 12-6 所示。其中有两个部分。

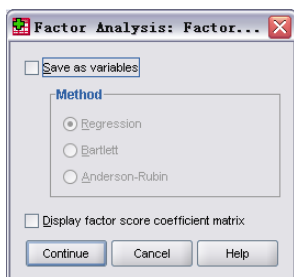


图 12-5 【Factor Scores】子对话框

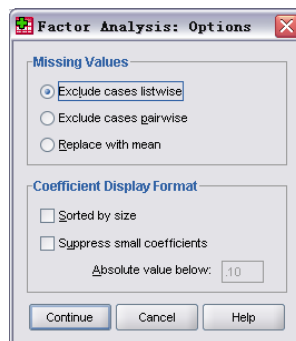


图 12-6 【Options】子对话框

- **Missing Values 单选框**

缺失值单选框，用于选择缺失值的处理方法。

- **Coefficient Display Format 复选框**

系数显示格式复选框，用于选择系数的输出方式。第一项是指将因子载荷矩阵和结构矩阵按数值大小排序；第二项是指不显示那些绝对值小于指定数值的载荷系数，在其后的栏内输入这个指定数值，要求取值范围为 0 到 1 之间，系统默认参数是 0.1。

12.1.3 引例及结果解释

下面通过两个例子来介绍【Factor Analysis】过程的操作及其结果。

例 12.1 以数据文件“各地区城市市政设施.sav”为例，其中的变量包括“city”——

城市名称、“X1”——年末实有道路长度（公里）、“X2”——年末实有道路面积（万平方米）、“X3”——城市桥梁（座）、“X4”——城市排水管道长度（公里）、“X5”——城市污水日处理能力（万立方米）、“X6”——城市路灯（盏）。数据如表 12-1 所示。（数据来源：《中国统计年鉴》—2005）

表 12-1 各地区城市市政设施表

city	X1	X2	X3	X4	X5	X6
北京	7482.7	11212.5	1285	6790	272	256032
天津	4240.3	5897.2	511	9332	93	181072
河北	7996.3	14987.7	1271	9575	279	321439
山西	4562.1	6471.8	752	3114	116	259914
内蒙古	3627.8	5935.9	278	4032	101	376329
辽宁	10407.3	15635.3	1300	9308	422	664359
吉林	4563.4	7165.8	451	4817	136	213881
黑龙江	9096.4	10731.3	656	5739	249	428561
上海	11028.0	19795.0	7297	6469	453	267442
江苏	26597.9	35596.2	12680	25538	1018	1169011
浙江	11288.7	18776.8	5847	16942	504	642965
安徽	7262.9	12109.1	1047	6680	307	264264
福建	4643.7	6801.7	1231	5427	196	290098
江西	3670.8	6071.6	428	3224	113	324801
山东	23617.0	40082.8	3712	20083	510	662650
河南	6505.5	13828.8	1027	8623	250	397351
湖北	14434.1	19958.9	1832	8791	426	303367
湖南	5539.9	8788.1	504	4946	328	255498
广东	22528.6	38856.0	3712	25168	543	1108886
广西	4761.0	7272.5	548	3774	282	332056
海南	1096.6	2234.2	126	1878	41	83849
重庆	3448.4	5206.1	630	3753	63	179468
四川	8263.8	14015.4	1926	8947	203	642540
贵州	2057.9	2623.0	300	3184	23	100437
云南	2502.6	3393.3	517	2653	161	162611
西藏	407.9	429.0	32	220		11085
陕西	3060.5	5526.7	394	2919	41	156488
甘肃	2810.2	4813.3	307	2620	71	118703
青海	539.9	888.7	63	535	9	22856
宁夏	1215.1	2317.6	120	861	54	118508
新疆	3706.4	5532.4	308	2940	124	215017

利用因子分析过程，分析各个城市的市政设施建设情况。执行以下操作：

执行【Analyze】/【Dimension Reduction】/【Factor】命令，弹出【Factor Analysis】

对话框	
Variables : X1、X2、X3、X4、X5、X6	选入分析变量
单击【Descriptives】按钮	弹出【Descriptives】子对话框
勾选 Coefficients	输出相关系数矩阵
KMO and Bartlett's test of sphericity	进行因子分析适用条件的检验
单击【Continue】按钮	回到主对话框
单击【Extraction】按钮	弹出【Extraction】子对话框
勾选 Display : Sree plot	输出碎石图
单击【Continue】按钮	回到主对话框
单击【Scores】按钮	弹出【Factor Scores】子对话框
勾选 Display factor score coefficient matrix	输出因子得分系数矩阵
单击【Continue】按钮	回到主对话框
单击【Options】按钮	弹出【Options】子对话框
选择 Replace with mean	用均值代替缺失值
单击【Continue】按钮	回到主对话框
单击【OK】按钮	生成以下结果

表 12-2 是 6 个分析变量的相关系数矩阵表,从表中可以看出这 6 个变量具有高相关性。

表 12-2 相关系数矩阵表

Correlation Matrix							
	年末实有 道路长度 (公里)	年末实有 道路面积 (万平方米)	城市桥梁 (座)	城市排水 管道长度 (公里)	城市污水 日处理能力 (万立方米)	城市路灯 (盏)	
Correlation 年末实有道路长度 (公里)	1.000	.983	.783	.939	.896	.883	
年末实有道路面积 (万平方米)	.983	1.000	.738	.940	.853	.867	
城市桥梁 (座)	.783	.738	1.000	.759	.873	.719	
城市排水管道长度 (公里)	.939	.940	.759	1.000	.845	.916	
城市污水日处理能 力 (万立方米)	.896	.853	.873	.845	1.000	.822	
城市路灯 (盏)	.883	.867	.719	.916	.822	1.000	

表 12-3 是 KMO 检验和 Bartlett 球形检验结果表。KMO 检验用于检验变量间的偏相关系数是否过小,一般情况下,当 KMO 大于 0.9 时效果最佳,小于 0.5 时不适宜做因子分析。Bartlett 球形检验用于检验相关系数矩阵是否是单位阵,如果结论是不拒绝该假设,则表示各个变量都是各自独立的。从表 12-3 可以看到 KMO 检验结果为 0.856,接近 0.9,很适合做因子分析,Bartlett 球形检验的 Sig.取值 0.000,表示拒绝该假设,认为各个变量之间不是独立的。

表 12-3 KMO 检验和 Bartlett 球形检验结果表

KMO and Bartlett's Test		
Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		.856
Bartlett's Test of Sphericity	Approx. Chi-Square	281.248
	df	15
	Sig.	.000

表 12-4 是变量共同度表，表中给出了提取公共因子前后各变量的共同度，它是衡量公共因子的相对重要性指标。比如表格的第一行数据说明变量“X1”的共同度为 0.954，即提取的公共因子对变量“X1”的方差做出了 95.4% 的贡献。通俗地说，就是指变量“X1”中 95.4% 的信息已经被提取出来。

表 12-4 变量共同度表

Communalities		
	initial	Extraction
年末实有道路长度 (公里)	1.000	.954
年末实有道路面积 (万平方米)	1.000	.919
城市桥梁 (座)	1.000	.742
城市排水管道长度 (公里)	1.000	.924
城市污水日处理能 力 (万立方米)	1.000	.882
城市路灯 (盏)	1.000	.859

Extraction Method: Principal Component Analysis

表 12-5 是主成分列表，表中列出了所有的主成分，且按照特征根从大到小次序排列。从表中可见，第一主成分特征根为 5.280，方差贡献率为 88.001%，前两个主成分的累计贡献率为 94.504%，根据提取因子的条件——特征值大于 1，本例只选出了一个因子。

表 12-5 主成分表

Total Variance Explained						
Component	Initial Eigenvalues			Extraction Sums of Squared Loadings		
	Total	% of Variance	Cumulative%	Total	% of Variance	Cumulative%
1	5.280	88.001	88.001	5.280	88.001	88.001
2	.390	6.503	94.504			
3	.162	2.707	97.211			
4	.104	1.738	98.950			
5	.051	.849	99.799			
6	.012	.201	100.000			

Extraction Method: Principal Component Analysis.

图 12-7 是碎石图，是按照特征根大小排列的主成分散点图。图中纵坐标为特征值，横坐标为因子数。从图中可见，除第一个主成分以外，其他的主成分特征根都很低。

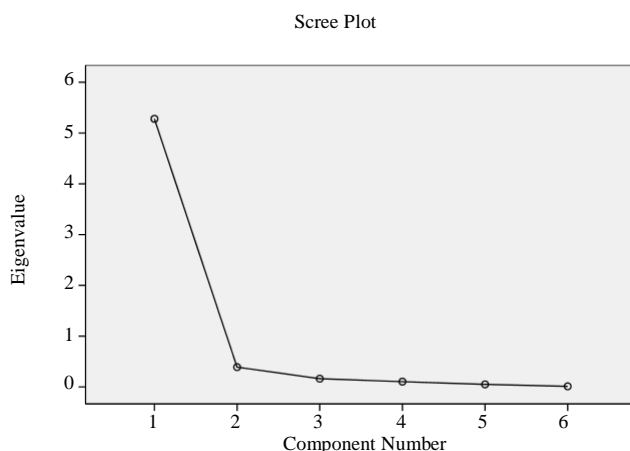


图 12-7 碎石图

表 12-6 为因子负荷矩阵，用来反映各个变量的变异可以主要由哪些因子解释。通过这个矩阵就可以给出各变量的因子表达式

$$X_1 = 0.977F_1 + \varepsilon_1, \dots, X_6 = 0.927F_1 + \varepsilon_6$$

因为只提取了一个公共因子，所以表达式中含有特殊因子。

表 12-7 是因子得分系数矩阵。通过此表就可以得出用各个变量的线性组合表达的主成分。本例的表达式就是

$$F_1 = 0.185X_1 + 0.182X_2 + 0.163X_3 + 0.182X_4 + 0.178X_5 + 0.176X_6$$

表 12-8 是因子得分的协方差矩阵，用来反映各因子间的联系程度。本例中只提取出了一个公共因子，故表格内容无实际意义。

表 12-6 因子负荷矩阵

Component Matrix ^a	
	Component
	1
年末实有道路长度 (公里)	.977
年末实有道路面积 (万平方米)	.959
城市桥梁 (座)	.862
城市排水管道长度 (公里)	.961
城市污水日处理能 力 (万立方米)	.939
城市路灯 (盏)	.927

Extraction Method: Principal Component Analysis.

a. 1 components extracted.

表 12-7 因子得分系数矩阵

Component Score Coefficient Matrix	
	Component
	1
年末实有道路长度 (公里)	.185
年末实有道路面积 (万平方米)	.182
城市桥梁 (座)	.163
城市排水管道长度 (公里)	.182
城市污水日处理能 力 (万立方米)	.178
城市路灯 (盏)	.176

Extraction Method: Principal Component Analysis.

表 12-8 因子得分的协方差矩阵

Component Score Covariance Matrix	
Component	1
1	1.000

Extraction Method: Principal Component Analysis.

下面再来看一个因子旋转的例子。

例 12.2 以数据文件“主要城市日照数.sav”为例，其中的变量包括城市的名称“city”、各个月份的日照数“Jan”、“Feb”、…、“Dec”。数据见第 11 章的表 11-10。（数据来源：《中国统计年鉴》—2005）。

利用因子分析过程分析这一年内各个城市的日照情况。执行以下操作：

执行【Analyze】/【Dimension Reduction】/【Factor】命令，弹出【Factor Analysis】对话框	
Variables : Jan、Feb、Mar、……、Dec	选入分析变量
单击【Descriptives】按钮	弹出【Descriptives】子对话框
勾选 KMO and Bartlett ' s test of sphericity	进行因子分析适用条件的检验
单击【Continue】按钮	回到主对话框
单击【Extraction】按钮	弹出【Extraction】子对话框
勾选 Display : Sree plot	输出碎石图
单击【Continue】按钮	回到主对话框
单击【Scores】按钮	弹出【Factor Scores】子对话框
勾选 Display factor score coefficient matrix	输出因子得分系数阵
单击【Continue】按钮	回到主对话框

单击【Rotation】按钮	弹出【Rotation】子对话框
选择 Varimax	采用方差最大化正交旋转
勾选 Loading plots	输出因子载荷图
单击【Continue】按钮	回到主对话框
单击【OK】按钮	生成以下结果

表 12-9 所示是 KMO 检验和 Bartlett 球形检验结果表。从表中可以看出 KMO 检验结果为 0.798，大于 0.5，说明本例比较适合做因子分析，Bartlett 球形检验的 Sig.取值 0.000，表示拒绝各变量是独立的假设。

表 12-9 KMO 检验和 Bartlett 球形检验结果表

KMO and Bartlett's Test		
Kaiser-Meyer Olkin Measure of Sampling Adequacy.		.798
Bartlett's Test of Sphericity	Approx. Chi-Square	437.331
	df	66
	Sig.	.000

表 12-10 是变量共同度表。表格的第一行数据说明变量“Jan”的共同度为 0.915，即选取的公共因子提取了变量“Jan”91.5%的信息。

表 12-10 变量共同度表

Communalities		
	Initial	Extraction
一月日照时数	1.000	.915
二月日照时数	1.000	.918
三月日照时数	1.000	.896
四月日照时数	1.000	.933
五月日照时数	1.000	.882
六月日照时数	1.000	.778
七月日照时数	1.000	.617
八月日照时数	1.000	.874
九月日照时数	1.000	.754
十月日照时数	1.000	.863
十一月日照时数	1.000	.847
十二月日照时数	1.000	.854

Extraction Method: Principal Component Analysis.

表 12-11 是主成分列表。从表中可见，第一主成分特征根为 6.854，方差贡献率为 57.041%，前 3 个主成分的累计贡献率为 84.421%，根据提取因子的条件——特征值大于 1，本例选出了 3 个因子。

表 12-11 主成分列表

Component	Initial Eigenvalues			Extraction Sums of Squared Loadings			Rotation Sums of Squared Loadings		
	Total	% of Variance	Cumulative%	Total	% of Variance	Cumulative%	Total	% of Variance	Cumulative%
1	6.845	57.041	57.041	6.845	57.041	57.041	4.581	38.173	38.173
2	1.962	16.347	73.388	1.962	16.347	73.388	2.886	24.047	62.220
3	1.324	11.034	84.421	1.324	11.034	84.421	2.664	22.201	84.421
4	.725	6.045	90.466						
5	.394	3.283	93.749						
6	.250	2.085	95.833						
7	.171	1.423	97.256						
8	.104	.870	98.126						
9	.080	.670	98.796						
10	.065	.539	99.335						
11	.047	.395	99.731						
12	.032	.269	100.000						

Extraction Method: Principal Component Analysis.

图 12-8 是碎石图，就是按照特征根大小排列的主成分散点图。从图中可见，前 3 个主成分的特征根都在 1 以上。

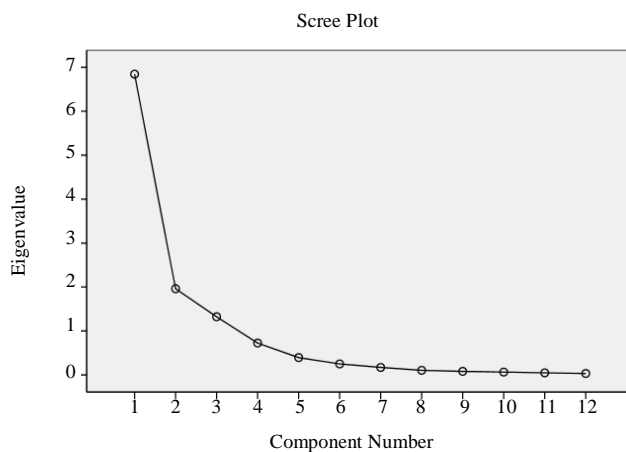


图 12-8 碎石图

表 12-12 所示为因子负荷矩阵。

通过这个矩阵就可以给出各变量的因子表达式：

$$X_1 = 0.852F_1 - 0.435F_2 - 0.015F_3$$

$$X_2 = 0.854F_1 - 0.419F_2 - 0.115F_3$$

...

$$X_{12} = 0.562F_1 - 0.164F_2 + 0.715F_3$$

表 12-12 因子载荷矩阵

	Component Matrix ^a		
	Component		
	1	2	3
一月日照时数	.852	-.435	-.015
二月日照时数	.854	-.419	-.115
三月日照时数	.869	-.275	-.257
四月日照时数	.805	-.079	-.528
五月日照时数	.888	-.033	-.303
六月日照时数	.764	.439	-.038
七月日照时数	.364	.644	-.265
八月日照时数	.465	.809	.066
九月日照时数	.794	.295	.192
十月日照时数	.800	.251	.400
十一月日照时数	.825	-.275	.300
十二月日照时数	.562	-.164	.715

Extraction Method: Principal Component Analysis.

a. 3 components extracted.

表 12-13 是经过正交旋转过后的因子载荷矩阵。

通过这个矩阵就可以给出旋转后的各变量的因子表达式：

$$X1 = 0.837F1' - 0.014F2' + 0.463F3'$$

$$X2 = 0.882F1' + 0.013F2' + 0.375F3'$$

...

$$X12 = 0.140F1' + 0.018F2' + 0.913F3'$$

并且从表中可以看出，第一主因子主要由前 5 个变量决定，第二主因子主要由中间 4 个变量决定，第三主因子主要由后 3 个变量决定。

表 12-13 旋转后的因子载荷矩阵

	Rotated Component Matrix ^a		
	Component		
	1	2	3
一月日照时数	.837	-.014	.463
二月日照时数	.882	.013	.375
三月日照时数	.901	.163	.241
四月日照时数	.903	.340	-.049
五月日照时数	.834	.392	.179
六月日照时数	.405	.730	.285
七月日照时数	.128	.763	-.134
八月日照时数	-.031	.917	.178
九月日照时数	.376	.588	.516
十月日照时数	.297	.528	.704
十一月日照时数	.592	.081	.700
十二月日照时数	.140	.018	.913

Extraction Method: Principal Component Analysis.

Rotation Method: Varimax with Kaiser Normalization.

a. Rotation converged in 4 iterations.

表 12-14 所示是因子转换矩阵。

表 12-14 因子转换矩阵

Component Transformation Matrix

Component	1	2	3
1	.754	.437	.491
2	-.432	.892	-.131
3	-.495	-.113	.861

Extraction Method: Principal Component Analysis.

Rotation Method: Varimax with Kaiser Normalization.

旋转前的因子载荷矩阵乘以因子转换矩阵就等于旋转后的因子载荷矩阵。

图 12-9 所示是因子旋转前的因子载荷图，图 12-10 是因子旋转后的因子载荷图（这里是为了做对比把这个图放在这里，实际的输出中只有图 12-10）。图中的坐标轴就是各个主因子，两个图中显示的分别是旋转前和旋转后原变量的位置。

表 12-15 是因子得分系数矩阵。从表中得到因子得分表达式，如下所示：

$$F1' = 0.195X1 + 0.229X2 + 0.252X3 + \cdots - 0.169X12$$

$$F2' = -0.142X1 - 0.126X2 - 0.048X3 + \cdots - 0.100X12$$

$$F3' = 0.081X1 + 0.015X2 - 0.086X3 + \cdots + 0.516X12$$

Component Plot

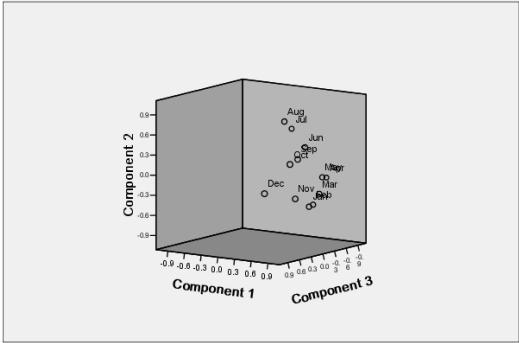


图 12-9 因子旋转前的因子载荷图

Component Plot in Rotated Space

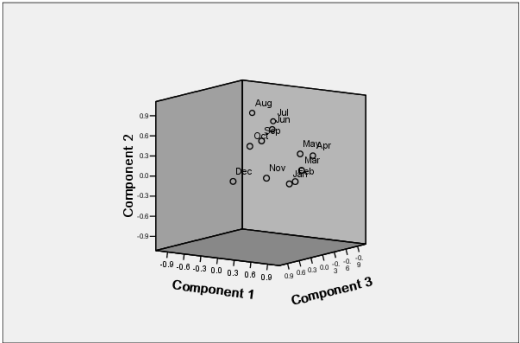


图 12-10 因子旋转后的因子载荷图

表 12-15 因子得分系数矩阵

Component Score Coefficient Matrix

	Component		
	1	2	3
一月日照时数	.195	-.142	.081
二月日照时数	.229	-.126	.015
三月日照时数	.252	-.048	-.086
四月日照时数	.304	.060	-.280
五月日照时数	.218	.067	-.131
六月日照时数	.002	.252	.001
七月日照时数	-.002	.339	-.189

续表

	Component		
	1	2	3
八月日照时数	-.151	.392	.022
九月日照时数	-.049	.169	.162
十月日照时数	-.117	.131	.301
十一月日照时数	.039	-.098	.273
十二月日照时数	-.169	-.100	.516

Extraction Method: Principal Component Analysis.

Rotation Method: Varimax with Kaiser Normalization.

表 12-16 是因子得分的协方差矩阵。由于因子得分的协方差矩阵为单位矩阵，说明提取的 3 个公共因子之间是不相关的。

表 12-16 因子得分的协方差矩阵

Component Score Covariance Matrix

Component	1	2	3
1	1.000	.000	.000
2	.000	1.000	.000
3	.000	.000	1.000

Extraction Method: Principal Component Analysis.

Rotation Method: Varimax with Kaiser Normalization.

12.2 简单对应分析——Correspondence Analysis过程

对应分析在 SPSS 中被分为简单对应分析和多元对应分析，分别被放在【Correspondence Analysis】过程和【Optimal Scaling】过程中。这一节就先来介绍【Correspondence Analysis】过程，它主要用于研究两个分类变量之间的关系。它在社会调查和专家评议调查中使用最为广泛。

12.2.1 简单对应分析基本原理

对应分析实际上是在 R 型分析和 Q 型分析基础上发展起来的一种多元统计方法。它将 R 型分析和 Q 型分析结合起来进行统计分析。另外根据两种分析的内在联系，可以将变量和样品同时反映到相同坐标轴（因子轴）的一张图形上，便于对问题进行分析。

简单对应分析又称为列联表对应分析，是对两个定性变量的多种状态进行对应性研究。对应分析法依靠主成分分析中的降维手段，可以更直观、明了地观察和分析定性变量在多种状态间的相互关系。

进行对应分析，首先要建立列联表（又叫对应分析表），表中的元素 k_{ij} 表示属于第一个变量的状态 i ，同时又属于第二个变量的状态 j 的样本点的个数。然后对列联表进行行剖面和列剖面的处理，这是很重要的一步。接下来，可以对列联表做统计分析，第一步是要判断两个定性变量的联系，一般采用卡方检验来验证两个变量独立的显著性。最后分别对行剖面点集和列剖面点集做主成分分析，以判断两个变量的联系情况。

12.2.2 Correspondence Analysis过程界面操作介绍

执行【Analyze】/【Dimension Reduction】/【Correspondence Analysis】命令，弹出如图 12-11 所示的【Correspondence Analysis】（简单对应分析）对话框。

1. 【Row】框

行变量框，用于放置行变量。单击下面的【Define Range】按钮，弹出如图 12-12 所示的【Define Row Range】（定义行变量取值范围）子对话框。其中包含两个部分。

- Category range for row variable 框

为行变量定义最大、最小值，输入相应的数值后，单击【Update】按钮，确认输入。

- Category Constraints 框

分类限制框。定义好变量取值范围后，类别的番号出现在此框中。旁边有 3 个单选项，用于对这些类别进行更进一步的设置。第一个选项 None 指无任何设置；第二项是指将某些分类合并，即令这些分类得分相同；第三项是指设置某些分类为追加分类，这些分类不进入分析，但是其单元格频数分布情况将显示在感知图中。

2. 【Column】框

列变量框，用于放置列变量。单击下面的【Define Range】按钮，弹出【Define Column Range】（定义列变量取值范围）子对话框。这个子对话框与图 12-12 显示的子对话框一样。

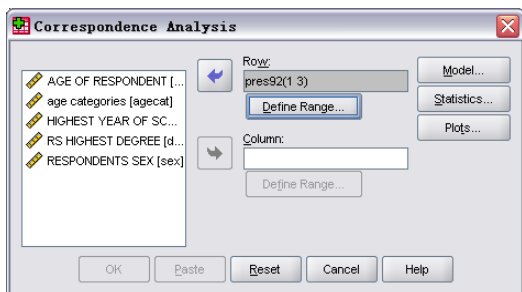


图 12-11 【简单对应分析】对话框

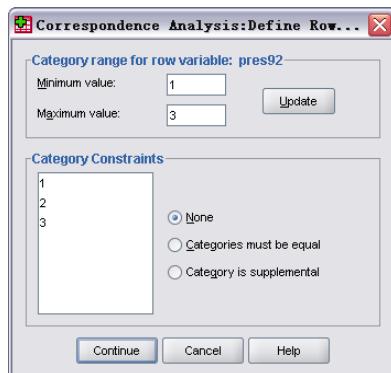


图 12-12 【Define Row Range】子对话框

3. 【Model】子对话框

模式子对话框。单击【Model】按钮，弹出此对话框，如图 12-13 所示。其中包含 4 个部分。

- Dimension in solution 栏

解维数栏，用于设置分析结果维数。在栏内输入一个小于各变量中的最少分类数的正整数，系统默认值为 2。

- Distance Measure 单选框

选择距离测量方式。包含两个选项，一个是 Chi square（卡方距离），它主要用于分类变量；另一个是 Euclidean（欧氏距离），主要用于连续变量，系统默认为卡方距离。

- **Standardization Method** 单选框

选择变量的标准化方式。其中的部分选项只有选择了欧氏距离后才能被激活。

- **Normalization Method** 单选框

选择正则化方法。系统默认为 Symmetrical（对称法）。

4. 【Statistics】子对话框

统计子对话框，设置需要输出的统计量。单击【Statistics】按钮，弹出此对话框，如图 12-14 所示。其中各选项的含义如下。

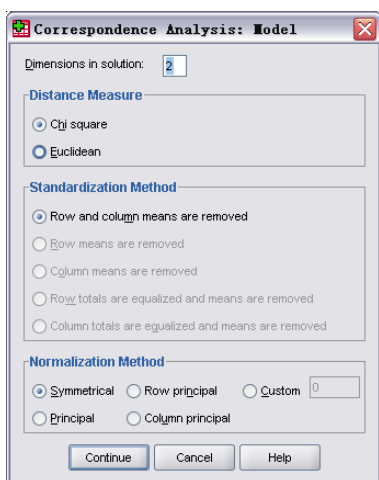


图 12-13 【Model】子对话框

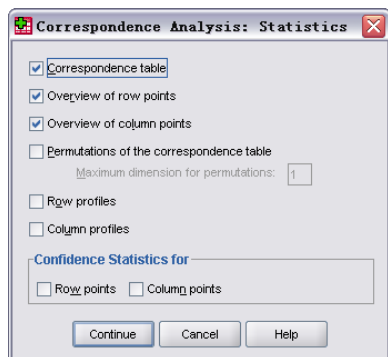


图 12-14 【Statistics】子对话框

- ① **Correspondence table**: 输出对应分析表，实际上就是列联表。
- ② **Overview of row points**: 输出行点概述表，将在结果分析中介绍。
- ③ **Overview of column points**: 输出列点概述表。
- ④ **Row profiles**: 输出行轮廓表。该表格按行变量的不同取值列出每个列变量所占的百分比。
- ⑤ **Column profiles**: 输出列轮廓表。该表格按列变量的不同取值列出每个行变量所占的百分比。
- ⑥ **Permutations of the correspondence table**: 为指定的前 n 个维度输出基于行列得分的原始表格。在下面的 Maximum dimension for permutations 栏内输入这个指定数目 n 。
- **Confidence Statistics** 复选框: 输出行点和列点的标准差，以及各维度坐标间的相关系数。

5. 【Plots】子对话框

图形子对话框，设置需要输出图形。单击【Options】按钮，弹出此对话框，如图 12-15 所示。其中包含 3 个部分。

- ① **Scatterplots** 复选框: 输出对应分析图。其中的选项包括 Biplot（双变量散点图）、Row points（行点图）和 Column points（列点图）。在下面的 ID label width for 栏内输入一个数值，用于限制标签长度，系统默认为 20。

② Line plots 复选框:输出各行变量分类对应于行得分的散点图,即第一个选项;输出各列变量分类对应于列得分的散点图,即第二个选项。

③ Plot Dimensions 单选框:图形维度单选框。可以选择输出分析结果的所有维度或者限制输出维度的数目。

12.2.3 引例及结果分析

下面通过一个例子介绍【Correspondence Analysis】过程的操作及其结果。

例 12.3 以数据文件“voter.sav”为例,此文件中的变量包括“pres92”——总统候选人、“age”——年龄、“agecat”——年龄类别、“educ”——教育年数、“degree”——学历水平、“sex”——性别。

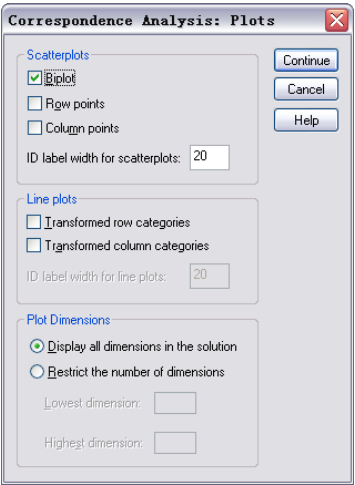


图 12-15 【Plots】子对话框

利用简单因子分析过程分析不同年龄段选民的倾向。执行以下操作:

执行【Analyze】/【Dimension Reduction】/【Correspondence Analysis】命令,弹出对话框	
Row : pres92	选入行变量
单击【Define Range】按钮	弹出【Define Row Range】子对话框
Minimum value : 1	
Maximum value : 3	定义行变量取值范围为 1 ~ 3
单击【Update】按钮	
单击【Continue】按钮	回到主对话框
Column : agecat	选入列变量
单击【Define Range】按钮	
Minimum value : 1	
Maximum value : 4	定义列变量取值范围为 1 ~ 4
单击【Update】按钮	
单击【Continue】按钮	回到主对话框
单击【OK】按钮	生成以下结果

结果浏览窗口中首先出现该对应分析模块的版权信息。

表 12-17 是对应分析表,即列联表。从图中可以看出不同年龄阶段的人群分别投票给三位总统候选人的人数。

表 12-18 是结果汇总表。表中列出了 Dimension(维数)、Singular Value(奇异值)、Inertia(惯量)、Chi Square(总的卡方检验)及 Sig.的值。其中,惯量指特征根,用于说明对应分析各个维度的结果能够解释列联表中两变量联系的程度;奇异值指惯量的平方根。表中的

两个维度分别解释了总信息量的 99.6% 和 0.4%，说明二维图形完全可以表示两个变量间的信息，且观察时以第一维度为主。

表 12-17 对应分析表

Correspondence Table

VOTE FOR CLINTON, BUSH, PEROT	age categories				
	It 35	35-34	45-64	65 +	Active Margin
Bush	153	156	219	133	661
Perot	99	81	82	16	278
Clinton	186	207	316	199	908
Active Margin	438	444	617	348	1847

表 12-18 结果汇总表

Summary

Dimension	Singular Value	Inertia	Chi Square	Sig.	Proportion of Inertia		Confidence Singular Value	
					Accounted for	Cumulative	Standard Deviation	Correlation
								2
1	.175	.030			.996	.996	.020	-.097
2	.011	.000			.004	1.000	.023	
Total		.031	56.531	.000 ^a	1.000	1.000		

a. 6 degrees of freedom

表 12-19 是行点概述表。表中给出行变量的 3 个分组在两个维度中的分值，其中 **Mass** 项指每一组所占的百分比，后面的 1、2 两项分别为分组在第一维度和第二维度的坐标值，右侧的 **Contribution** 项给出了每个分组对各个维度的贡献量，包括点对维度惯量的贡献和维度对点惯量的贡献。

表 12-19 行点概述表

Overview Row Points^a

VOTE FOR CLINTON, BUSH, PEROT	Mass	Score in Dimension		Inertia	Contribution				
		1	2		of Point to Inertia of Dimension		of Dimension to Inertia of Point		
					1	2	1	2	Total
Bush	.358	.072	-.137	.000	.011	.631	.820	.180	1.000
Perot	.151	.975	.046	.025	.819	.030	1.000	.000	1.000
Clinton	.492	.246	.085	.005	.170	.338	.993	.007	1.000
Active Total	1.000			.031	1.000	1.000			

a. Symmetrical normalization

表 12-20 是列点概述表。与行点概述表类似，这里不再赘述。

表 12-20 列点概述表

Overview Column Points ^a									
age categories	Mass	Score in Dimension		Inertia	Contribution				
					of Point to Inertia of Dimension		of Dimension to Inertia of Point		
					1	2	1	2	Total
It 35	.237	−.519	−.099	.011	.366	.218	.998	.002	1.000
35-44	.240	−.217	.019	.002	.065	.008	1.000	.000	1.000
45-64	.334	.126	.126	.001	.030	.502	.942	.058	1.000
65 +	.188	.707	−.124	.016	.539	.272	.998	.002	1.000
Active Total	1.000			.031	1.000	1.000			

a. Symmetrical normalization

图 12-16 为对应分析图。从这个图中可以看出所需要的分析结果。首先看同一变量的不同分组在某一侧维度上靠的远近程度，较近表示这些分组在该维度上区别不大，比如本例中的第二维度。然后看不同变量的各个分组间的位置，从图形的中心（0，0）点出发，相同方位上大致相同的区域内的不同变量的分组彼此有联系。按照这一原则，可以看出 45~64 岁这个阶段的选民都倾向于克林顿，其他的就没有很明显的倾向性了。

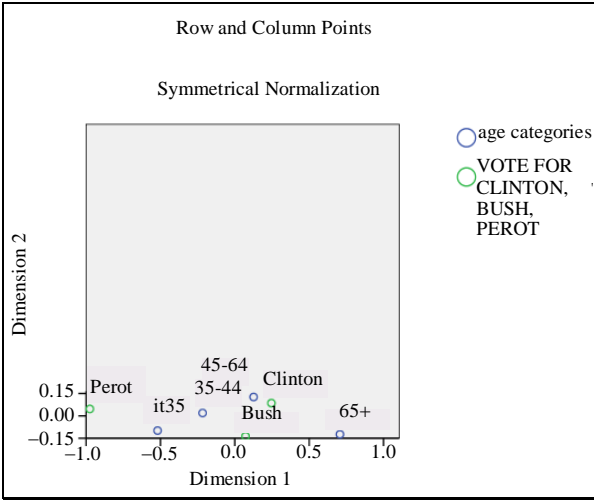


图 12-16 对应分析图

12.3 最优尺度分析——Optimal Scaling过程初步认识

SPSS 的【Optimal Scaling】过程实际上提供了三种分析方法，分别是 Multiple Correspondence Analysis（多元对应分析）、Categorical Principal Components（分类变量的主成分分析）和 Nonlinear Canonical Correlation（非线性典型相关分析），它们分别对应三种不同的模型，每个模型适用于不同的数据情况。

最优尺度分析的特点主要表现在：

- (1) 可以同时分析多个分类变量间的关系；
- (2) 可以处理各种类型的变量，如对无序多分类变量、有序多分类变量和连续性变量同时进行分析。
- (3) 对多选题的分析提供支持。

由于最优尺度分析法的缺点是不能自动筛选变量，因此变量较多时可能会影响分析结果。

执行【Analyze】/【Dimension Reduction】/【Optimal Scaling】命令，弹出如图 12-17 所示的【Optimal Scaling】（最优尺度）对话框。

1. 【Optimal Scaling】单选框

当分析变量中存在无序多分类变量时，则选择第二项，否则使用系统默认值。

2. 【Number of Sets of Variables】单选框

确定是在不同变量间进行分析还是在几组变量间进行分析。若数据中存在复选集变量，则应该选择第二项 Multiple sets。

3. 【Selected Analysis】框

显示分析方法框。如前所述，最优尺度过程提供了三种分析方法，这三种方法分别是。选定上面两个单选框中的选项后，系统自动选择适当的分析方法，被选中的分析方法加黑显示。下面分别介绍这 3 种方法。

• Multiple Correspondence Analysis

多元对应分析。所有变量均在名义测量时使用。该方法用于分析多个无序分类变量间的关系。多元对应分析方法与简单对应分析方法非常类似，只是分析的变量可以为多个。系统默认状态下，单击【Define】按钮，就弹出如图 12-18 所示的多元对应分析对话框。

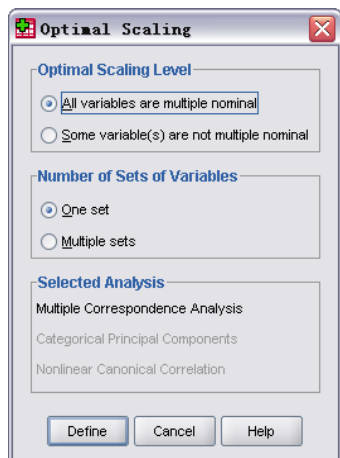


图 12-17 【最优尺度】对话框

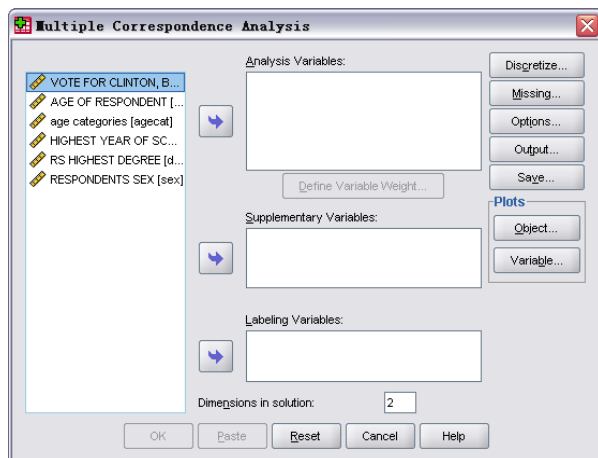


图 12-18 【多元对应分析】对话框

- Categorical Principal Components

分类变量的主成分分析。当一些变量为有序分类或者连续性变量时使用。该方法使用主成分提取方式，以尽量少的主成分解释尽量多的原始信息。在【Optimal Scaling】单选框中选择第二项，然后单击【Define】按钮，弹出分类变量的主成分分析对话框，其样式与多元对应分析对话框非常类似。

- Nonlinear Canonical Correlation

非线性典型相关分析。当分析变量中有复选集变量时，系统自动选择这种分析方法。该方法用于分析两个或者多个变量集之间的关系，允许变量为任何类型。在【Number of Sets of Variables】单选框中选择第二项，然后单击【Define】按钮，弹出如图 12-19 所示的多元对应分析对话框。

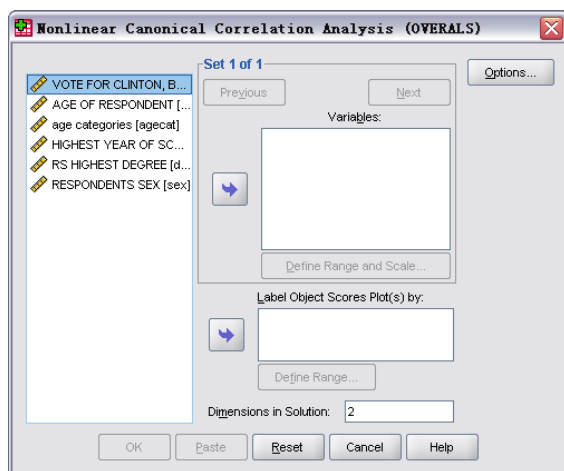


图 12-19 【多元对应分析】对话框

这三种分析方法的基本原理及其对应的对话框界面操作方法这里不再详细介绍了，感兴趣的读者可以查看 SPSS 自带的帮助文档。

12.4 本章小结

主成分分析、因子分析和对应分析是多元统计分析中非常重要的三类统计方法。本章介绍了它们在 SPSS 中的实现。其中，主成分分析在 SPSS 中没有单独的集成模块，而是和因子分析一起放在【Factor Analysis】过程中。

需要注意的是虽然主成分分析和因子分析都可以用【Factor Analysis】过程来实现。但是二者对观测量的数目要求有所不同。主成分分析对观测量没有严格要求，但是对应分析要求观测量至少为变量的五倍以上且越多越好。

同时，本章详细介绍了简单对应分析【Correspondence Analysis】过程，对于多元对应分析可以由【Optimal Scaling】过程实现，请读者自行研究学习。

第 13 章 非参数检验

前面介绍过许多分析过程都是涉及参数假设检验的，然而在很多实际问题中，往往并不知道总体所服从的分布，这时就需要根据观测资料来推断总体是否服从某种已知形式的分布，此时需要用到非参数检验。SPSS 为非参数检验专门提供了一个子菜单——【Nonparametric Tests】子菜单，本章通过例子学习非参数检验及其在 SPSS 中的实现。本章内容包括：

- 非参数检验相关原理简介
- 分布类型的检验
- 分布位置检验

13.1 非参数检验相关原理简介

本节将概括性地介绍非参数检验的基本概念、优缺点和类型。对于各类具体的非参数检验方法，将在本章的后续几节给出详细地介绍。

13.1.1 非参数检验的概念

非参数检验与参数检验是两个相对的概念，先来说说参数检验。一般而言，一个典型的统计推断过程通常由 5 个步骤构成：假定分布族，抽样，计算统计量和抽样分布，进行推估和检验，评价模型。这里的第一步假定分布族是对实际问题的数学描述，它是统计推断的基础。比如，在研究保险公司的索赔请求数时，可能假定索赔请求数来自泊松分布 $p(\lambda)(0 < \lambda < \infty)$ 。样本被视为从分布族的某个参数族抽取出来的总体的代表，而未知的仅仅是总体分布具体的参数值，这样推断问题就转化为对分布族若干个未知参数的推断问题，用样本对这些参数做出估计或者进行某种形式的假设检验，将这类推断方法称为参数检验方法。

然而在许多实际问题中，人们往往对总体的分布形式知之甚少，很难对总体的分布形式和统计模型做出明确的假定。这种不假定总体分布的具体形式，尽量从数据（或样本）本身获得所需要的信息，通过推断方法获得结构关系，并逐步建立对事物的数学描述和统计模型的方法称为非参数检验方法。

因此简而言之，参数检验是已知总体分布，再来估计和检验参数的统计方法，而非参数检验则是在未知总体分布的情况下估计和检验总体分布的统计方法。两者最大的区别在

于是否已知总体分布。

13.1.2 非参数检验的优缺点

非参数统计学是统计学的一个分支。相对于参数统计而言,非参数统计有以下几个突出的特点。

(1) 由于非参数统计方法对总体的假定相对较少,因而有广泛的适用性,结果一般较好的稳定性,即不会产生由于总体分布的一些变化而导致的发生大的结论性错误。如果模型通不过检验,样本量不足是其中一个可能的原因,但是追加样本量在很多行业中代价是巨大的。另外一个潜在的原因则是模型假定本身存在问题,如果是后者,那么就可以通过改变方法,不是错误地采取增加样本量的方法达到分析目的。

(2) 非参数统计可以处理所有类型的数据。一般来说,参数统计主要是针对定量数据,如果所收集到的数据不符合参数统计模型的假定,则很多参数模型无能为力,此时只能尝试非参数方法。

(3) 容易计算。由于采用大样本原理,因而大部分非参数统计量都服从正态分布或是由正态分布导出的分布,很容易通过编写程序求解,计算结果也容易解释。

当然,非参数方法也有一些缺点,其中最大的缺点就是它的检验效能较低。

13.1.3 非参数检验的类型

对非参数统计方法的研究从 20 世纪 40~50 年代就已开始形成了,发展至今,非参数统计已经获得了长足的进步,很多非参数检验方法在实际问题中得到了成功地应用。按照不同的分类方法,非参数检验可以分为很多不同的类型。SPSS 为非参数检验提供了 8 种最基本、最简单的非参数检验方法,如图 13-1 所示。它们被分为两个大类——分布类型检验方法和分布位置检验方法。

1. 分布类型检验方法

检验样本所在总体是否服从已知的理论分布。图 13-1 所示的前 4 个子菜单都是属于分布类型检验方法。

2. 分布位置检验方法

检验样本所在总体的分布位置或者形状是否相同。图 13-1 所示的后 4 个子菜单都是属于分布位置检验方法。在实际问题中,分布位置检验方法有更广泛地应用。

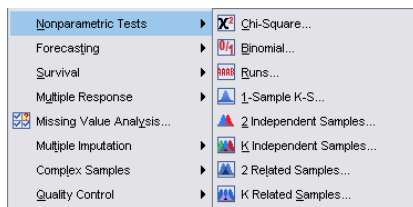


图 13-1 【Nonparametric Tests】子菜单

13.2 分布类型的检验

检验样本所在的总体是否服从已知的理论分布。SPSS 为这一类检验提供了 4 个子菜单，本节将通过例子详细介绍这 4 种检验方法的一般步骤、操作界面和结果解释。

13.2.1 卡方检验——Chi-Square过程

卡方检验法又称为卡方拟合优度检验，用于检验观测数据是否与某种概率分布的理论数值相符合，进而推断观测数据是否是来自于该分布的样本问题。

1. 卡方检验的统计原理

假设有一组分类数据，其中 n 个观测值可以分为 c 种类别，每一类别又可计算它的发生频率，称为实际观测频数，记为 $O_i (i=1, 2, \dots, c)$ 。这里想了解每一类别发生的概率是否与理论分布 $\{p_i : i=1, 2, \dots, c\}$ 一致。即有如下的假设检验问题：

H_0 ：总体分布服从理论分布；

H_1 ：总体分布不服从理论分布。

若 H_0 假设成立，则期望频数应为 $E_i = np_i (i=1, 2, \dots, c)$ ，因此设立假设检验问题如下：

由实际频数 (O_i) 与期望频数 (E_i) 是否接近作为检验总体分布与理论分布是否一致的测量标准，通常采用如下定义的卡方统计量：

$$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i} = \sum \frac{O_i^2}{E_i} - n$$

结论：当实际观测 χ^2 大于自由度 $v=c-1$ 的 χ^2 值，即 $\chi^2 < \chi_{\alpha, 1-c}^2$ ，（其中 α 为给定的显著性水平），则拒绝 H_0 ，表示数据分布与理论分布不符。

2. 卡方检验的数据要求及界面操作简介

卡方检验过程要求检验变量是否为数值型分类变量，且数据测度水平为 Ordinal 或 Nominal。若分类变量为字符型，可以使用【Transform】菜单下的【Recode】过程或者【Automatic Recode】过程将其转化为数值型变量，具体做法可参见本书的 3.3.3 节。若文件中变量为连续型变量，则可以使用【Transform】菜单下的【Recode】过程将样本空间划分区间或者分类。

执行【Analyze】/【Nonparametric Tests】/【Chi-Square】命令，弹出如图 13-2 所示的【Chi-Square Test】（卡方检验）对话框。下面介绍其中主要元素。

• 【Test Variable List】框

检验变量列表框，放置检验变量。可以选入多个检验变量到其中，但是系统进行处理时，是对每一个变量分别进行处理。

• 【Expected Range】单选框

期望检验范围框，用于选择检验范围。其中有两个选项，第一项是指检验范围是从原始数据最小值到最大值；第二项是指由用户自己指定一个检验范围，并分别在下面的【Lower】栏和【Upper】栏内输入这个指定范围的下限和上限。系统默认为第一项。

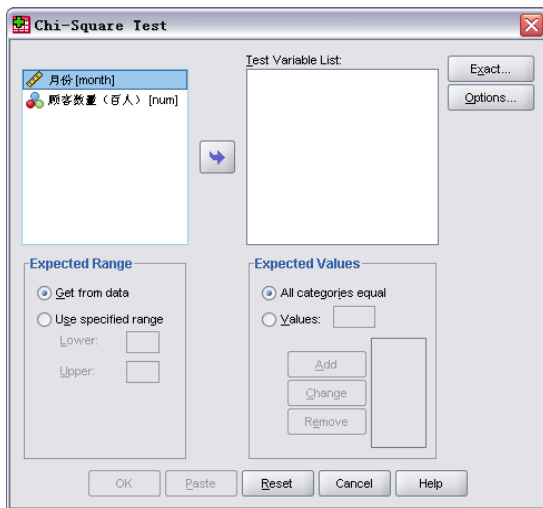


图 13-2 【卡方检验】对话框

- 【Expected Values】单选框

期望值单选框，用于输入各分类的期望值。不同的分布对每一个分类都有不同的期望值，【Chi-Square】过程就是利用不同的期望值来区分不同分类的。比如，若要检验变量值是否服从均匀分布，则选择第一项 All categories equal，表示每个分类的期望值都相同。而在其他分布中，每个分类的期望值是不同的，需要用户自己将计算好的期望值输入到第二个选项的【Values】栏内，可以直接输入期望概率值。但注意所有的期望概率值相加应该为 1，也可以输入整数值，系统会自动计算输入的每个整数值在所有输入的整数值中所占的比例，并将这个比例值视为所对应的每一类的期望概率值。

注意 输入期望值的顺序是非常重要的，它应该和检验变量的分类值的升序一一对应。

- 【Exact Tests】子对话框

精确检验子对话框，主要用于定义确切概率的计算。单击【Exact】按钮，弹出如图 13-3 所示的【Exact Tests】子对话框。其中包括三个单项。

- ① Asymptotic only 选项：只计算近似概率。
- ② Monte Carlo 选项：用 Monte Carlo 法计算精确概率。可自行设置置信度和抽样次数。
- ③ Exact 选项：在给定时间内计算精确概率的值，如果超出给定时间就停止计算。

- 【Options】子对话框

选项子对话框，用于选择统计量的输出和缺失值的处理方式。单击【Options】按钮，弹出如图 13-4 所示的【Options】子对话框。其中包括两个部分。

① Statistics 复选框：选择要输出的统计量。其中的选项有 Descriptive（描述统计量），表示输出样本的均值、标准差、最小值、最大值，以及非缺失的观测量数。选项 Quartiles（四分位数），表示输出变量对应于 25%、50% 和 75% 的四分位值，其中 50% 的四分位值就是指中位数。

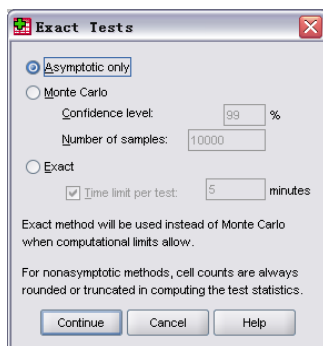


图 13-3 【Exact Tests】子对话框

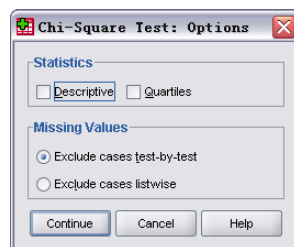


图 13-4 【Options】子对话框

② Missing Values 单选框：选择缺失值的处理方式，当有多个检验变量被选入时，此单选框才有意义。其中的第一个选项是指对每一个检验变量来个别地排除缺失值，第二个选项指凡含有缺失值的观测量全部从分析中排除。

3. 引例及结果解释

例 13.1 均匀分布的一致性检验。调查某美发店上半年各月顾客数量，调查结果如表 13-1 所示，检验各月顾客数是否服从均匀分布。

表 13-1 某美发店上半年各月顾客数量

月 份	1	2	3	4	5	6
顾客数量（百人）	27	18	15	24	36	30

假设检验问题：

H_0 ：各月顾客数符合均匀分布。

STEP 01 建立数据文件“某美发店上半年各月顾客数量.sav”，建立变量“month”——月份和变量“num”——顾客数量，输入以上数据。

STEP 02 将变量“num”定义为权变量。

执行【Data】/【Weight Cases】命令，弹出【Weight Cases】对话框

选择 Weight cases by

选择加权文件

Frequency Variable : num

选入权变量

单击【OK】按钮

定义完成

STEP 03 利用卡方检验来检验顾客人数是否服从均匀分布。

执行【Analyze】/【Nonparametric Tests】/【Chi-Square】命令，弹出对话框

Test Variable List : month

选入检验变量

单击【Options】按钮

进入【Options】子对话框

Statistics : 勾选 Descriptive

输出描述统计量

勾选 Quartiles

输出四分位数

单击【Continue】按钮

回到主对话框

单击【Ok】按钮

生成以下结果

表 13-2 所示为描述统计量表，表中依次列出了检验变量的观测量数目、均值、标准差、最小值、最大值，以及四分位数。

表 13-2 描述统计量表

Descriptive Statistics								
	N	Mean	Std. Deviation	Minimum	Maximum	Percentiles		
						25th	50th (Median)	75th
月份	150	3.76	1.790	1	6	2.00	4.00	5.00

表 13-3 是每个月份的顾客频数表。表中将每一个月份作为一类，显示了各类的实际观测量数目、期望观测量数，以及残差（即实际观测量数和期望观测量数之差）。

表 13-3 每个月份的顾客频数表

月份			
	Observed N	Expected N	Residual
1	27	25.0	2.0
2	18	25.0	-7.0
3	15	25.0	-10.0
4	24	25.0	-1.0
5	36	25.0	11.0
6	30	25.0	5.0
Total	150		

表 13-4 是卡方检验表，显示了卡方检验的结果。从表中可以看出，卡方统计量值为 12.000，自由度为 5，近似的显著性概率为 0.035，这个值小于 0.05，具有期望频数小于 5 的单元格为 0 个，最小期望的单元格频数为 25.0。

表 13-4 卡方检验表

Test Statistics	
	月 份
Chi-Square ^a	12.000
df	5
Asymp. Sig.	0.035

a. 0 cells(.0%) have expected frequencies less than

5. The minimum expected cell frequency is 25.0.

因此得出结论：拒绝零假设，认为到该店消费的顾客在各月比例不相等。

例 13.2 泊松分布的一致性检验。调查某农作物根部蚜虫的分布情况，调查结果如表 13-5 所示，问蚜虫在某农作物根部分布是否为泊松分布。

表 13-5 某农作物根部蚜虫分布情况表

每株虫数 x	0	1	2	3	4	□5
实际株数	10	24	10	4	1	1

假设检验问题：

H_0 ：蚜虫在农作物根部的分布是泊松分布。

若蚜虫在农作物根部的分布是泊松分布，则分布列为：

$$\hat{p}_x = P(X = x) = \frac{e^{-\lambda} \lambda^x}{x!}, x = 0, \dots, 5$$

其中 λ 是泊松分布的期望，是未知的，需要用观测量估计，其估值如下：

$$\hat{\lambda} = \bar{x} = (0 \times 10 + 1 \times 24 + \dots + 5 \times 1) / 50 = 1.3$$

由 $\hat{\lambda}$ 值计算得出每一类的期望概率值 $\hat{p}_0, \hat{p}_1, \hat{p}_2, \hat{p}_3, \hat{p}_4, \hat{p}_5$ 。

再利用卡方统计量判断检验结果。

操作步骤如下：

STEP 01 建立数据文件“某农作物根部蚜虫的分布情况.sav”，建立变量“x”——每株虫数和变量，“num”——实际株数，将以上数据依次输入。

STEP 02 将变量“num”定义为权变量。

执行【Data】/【Weight Cases】命令，弹出【Weight Cases】对话框

选择 Weight cases by 选择加权文件

Frequency Variable : num 选入权变量

单击【OK】按钮 定义完成

STEP 03 计算泊松分布的期望估计值 $\hat{\lambda}$ 。

执行【Analyze】/【Descriptives Statistics】/【Descriptives】命令，弹出对话框

Variables : x 选入分析变量

单击【OK】按钮 生成以下结果

表 13-6 为描述统计量表，表中列出了分析变量的观测量数目、最小值、最大值、均值和标准差。其中均值就是这里要求的期望估计值 $\hat{\lambda}$ ，从表中可以看出 $\hat{\lambda}=1.3$ 。

表 13-6 描述统计量表

Descriptive Statistics					
	N	Minimum	Maximum	Mean	Std. Deviation
每株虫数	50	0	5	1.30	1.074
Valid N (listwise)	50				

STEP 04 计算泊松分布的分布列（即每一类的期望概率值） $\hat{p}_0, \hat{p}_1, \hat{p}_2, \hat{p}_3, \hat{p}_4, \hat{p}_5$ 。

执行【Transform】/【Compute Variable】命令，弹出【Compute Variable】对话框

Target Variable : p 建立目标变量 p

Numeric Expression : 输入目标变量表达式

CDF.POISSON (x,1.3) -CDF.POISSON (x-1,1.3)

单击【OK】按钮

新变量 p 出现在数据文件中。

这里的变量 p 的值即每一类的期望概率值 $\hat{p}_0, \hat{p}_1, \hat{p}_2, \hat{p}_3, \hat{p}_4, \hat{p}_5$ 。

STEP 05 利用卡方检验检验蚜虫在农作物根部的分布是否是泊松分布。

执行【Analyze】/【Nonparametric Tests】/【Chi-Square】命令，弹出对话框

Test Variable List : x

选入检验变量

Expected Values :

Values : 2725、3543、2303

依次输入期望频数值

998、324、84

单击【Options】按钮

进入【Options】子对话框

Statistics : 勾选 Descriptive

输出描述统计量

勾选 Quartiles

输出四分位数

单击【Continue】按钮

回到主对话框

单击【OK】按钮

生成如下结果

表 13-7 为检验变量的描述统计表。前面已经介绍过，这里不再赘述。

表 13-7 描述统计表

Descriptive Statistics

	N	Mean	Std. Deviation	Minimum	Maximum	Percentiles		
						25th	50th (Median)	75th
每株虫数	50	1.30	1.074	0	5	1.00	1.00	2.00

表 13-8 为每株虫数的频数表。表中按照每株虫数的不同分类，共为了 5 类。

表 13-8 每株虫数的频数表

每株虫数

	Observed N	Expected N	Residual
0	10	13.7	-.37
1	24	17.8	6.2
2	10	11.5	-1.5
3	4	5.0	-1.0
4	1	1.6	-0.6
5	1	0.4	0.6
Total	50		

表 13-9 是卡方检验表，显示了卡方检验的结果。从表中可以看出，卡方统计量值为 4.617，自由度为 5，近似的显著性概率为 0.464，这个值远大于 0.05，具有期望频数小于 5 的单元格为两个，最小期望的单元格频数为 0.4。

表 13-9 卡方检验表

Test Statistics	
	每株虫数
Chi-Square ^a	4.617
df	5
Asymp.Sig.	.464

a. 2 cells (33.3%) have expected frequencies less than 5. The minimum expected cell frequency is 0.4.

因此得出结论：不能拒绝零假设，蚜虫在某作物根部的分布可能是泊松分布。

例 13.3 正态分布的一致性检验。从某地区高中二年级学生中随机抽取 45 名学生，量得体重如表 13-10 所示，问该地区学生体重（单位：公斤）的分布是否为正态分布。

表 13-10 某地区高二学生体重抽查结果

36	36	37	38	40	42	43	43	44	45	48	48	50	50	51
52	53	54	54	56	57	57	57	58	58	58	58	58	59	60
61	61	61	62	62	63	63	65	66	68	68	70	73	73	75

假设检验问题：

H_0 ：某地区高中二年级学生体重分布为正态分布。

STEP 01 建立数据文件“某地区高二学生体重抽查结果.sav”，建立变量“weight”一体重，将以上数据依次输入。

STEP 02 将上述体重分为 5 组。

执行【Transform】/【Recode Into Different Variables】命令，弹出对话框

Input Variable	Output Variable : weight	选入原变量
Output Variable : Name : interval		定义新变量名
Label : 区间		定义新变量标签
单击【Change】按钮		添加新变量名
单击【Old and New Values】按钮		打开【Old and New Values】子对话框
Old New:SYSMIS	SYSMIS	选择转化的方法
30 thru 40	1	
40 thru 50	2	
50 thru 60	3	
60 thru 70	4	
70 thru 80	5	
单击【Continue】按钮		回到主对话框
单击【OK】按钮		新变量“interval”出现在文件中

STEP 03 计算正态分布的分布参数估计值。

执行【Analyze】/【Descriptives Statistics】/【Descriptives】命令，弹出对话框

Variables : weight

选入分析变量

单击【OK】按钮

生成以下结果

表 13-11 为描述统计量表。通过此表可以得到正态分布的参数估计值： $\hat{\mu} = 55.36$ ，即体重的均值； $\hat{\sigma}^2 = 10.309$ ，即体重的标准差。

表 13-11 描述统计量表

Descriptive Statistics					
	N	Minimum	Maximum	Mean	Std Deviation
体重	45	36	75	55.36	10.309
Valid N (listwise)	45				

STEP 04 计算正态分布的期望概率。

先建立一个新变量“x”，其取值如图 13-5 所示，即每个区间的上限值。然后执行以下操作：

执行【Transform】/【Compute Variable】命令，弹出【Compute Variable】对话框

Target Variable : p

建立目标变量 p

Numeric Expression :

输入目标变量表达式

CDF.NORMAL (x,55.36,10.309) -CDF.NORMAL (LAG (x) ,55.36,10.309)

单击【OK】按钮

新变量 p 出现在数据文件中

图 13-5 中显示新变量“x”和“p”的取值。其中 p_i 的值是由计算得到的

$$P_1 = 1 - 0.2334 - 0.3721 - 0.2485 - 0.0694 = 0.0766$$

x	p
40	.0766
50	.2334
60	.3721
70	.2485
80	.0694

图 13-5 新变量 x 和 p 的取值

STEP 05 利用卡方检验检验这个地区高中二年级学生体重分布是否为正态分布。

执行【Analyze】/【Nonparametric Tests】/【Chi-Square】命令，弹出对话框

Test Variable List : interval

选入检验变量

Expected Values :

Values : 766、2334、2721

依次输入期望频数值

2485、694

单击【OK】按钮

生成以下结果

表 13-12 为每个区间的频数表。

表 13-12 每个区间的频数表

区间			
	Observed N	Expected N	Residual
1	5	3.4	1.6
2	9	10.5	-1.5
3	16	16.7	-0.7
4	12	11.2	-0.8
5	3	3.1	-0.1
Total	45		

表 13-13 是卡方检验表，显示了卡方检验的结果。从表中可以看出，卡方统计量值为 1.012，自由度为 4，近似的显著性概率为 0.908，这个值远远大于 0.05，具有期望频数小于 5 的单元格为两个，最小期望的单元格频数为 3.1。

表 13-13 卡方检验表

Test Statistics	
	区 间
Chi-Square ^a	1.012
df	4
Asymp. Sig.	0.908

a. 2 cells (40.0%) have expected frequencies less than
5. The minimum expected cell frequency is 3.1.

因此得出结论：不拒绝零假设，该地区高中二年级学生体重服从正态分布。

13.2.2 二项分布检验——Binomial过程

二项分布检验过程是用对二元变量的两个分类的观测频数与某个具有确定的概率参数的二项分布的期望频数进行比较的假设检验问题。

1. 二项分布检验的数据要求及界面操作简介

二项分布检验过程要求检验变量是数值型的二元变量。若变量不是二元变量，可以使用【Transform】菜单下的【Recode】过程将其数据分成两组，或者可以通过设置断点将数据分成两组。

执行【Analyze】/【Nonparametric Tests】/【Binomial】命令，弹出如图 13-6 所示的【Binomial Test】（二项分布检验）对话框。下面来介绍其中的主要元素。

• 【Test Variable List】框

检验变量列表框，放置检验变量。可以选入多个检验变量到检验变量列表框，但是系统进行处理时，是对每一个变量分别进行处理。

• 【Define Dichotomy】单选框

定义二元变量单选框。当检验变量已经是二元变量时，选择第一项 Get from data，这是系统默认选项；当检验变量不是二元变量时，选择第二项 Cut point（断点），并在其后

的空白栏内输入断点值，输入好断点值以后，系统自动将变量值小于或等于断点值的观测量分为一组，其余的分为另一组。

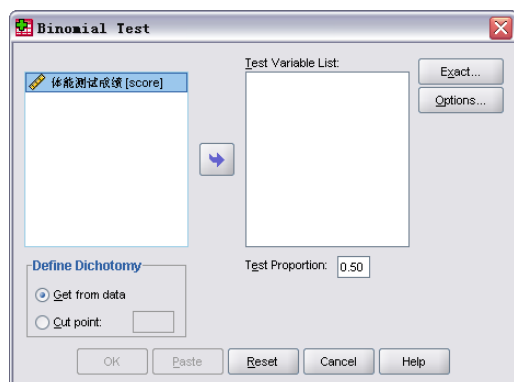


图 13-6 【二项分布检验】对话框

• 【Test Proportion】栏

检验比率栏，指定检验的零假设。在其后的栏内输入一个范围在 0.001~0.999 之间的数值，作为待检验的第一组的概率。系统默认值为 0.5，若使用系统默认值，则结果中输出双尾检验结果，否则输出单尾检验结果。

• 【Exact Tests】子对话框

单击【Exact】按钮，弹出【Exact Tests】子对话框。

• 【Options】子对话框

单击【Options】按钮，弹出【Options】子对话框。

上面两个子对话框与图 13-3 和图 13-4 所示的子对话框一致。

2. 引例及结果解释

例 13.4 二项分布的一致性检验。表 13-14 是 16 名学生在了一项体能测试上的成绩，以 60 分作为及格线，学校要求及格率达到 90%，问根据这批抽样数据，体能及格率是否达到了 90%？

表 13-14 体能测试成绩抽样结果

82	53	70	73	103	71	69	80
54	38	87	91	62	75	65	77

假设检验问题：

H_0 : 体能测试及格率达到了 90%。

STEP 01 建立数据文件“体能测试成绩抽样.sav”，建立变量“score”——体能测试成绩，将以上数据依次输入。

STEP 02 利用二项分布检验检验体能测试及格率是否达到 90%。

执行【Analyze】/【Nonparametric Tests】/【Binomial】命令，弹出【Binomial Test】对话框

Test Variable List : score	选入检验变量
Define Dichotomy : Cut point 60	选择断点 60
单击【Options】按钮	进入【Options】子对话框
Statistics : 勾选 Descriptive	输出描述统计量
勾选 Quartiles	输出四分位数
单击【Continue】按钮	回到主对话框
单击【Ok】按钮	生成以下结果

表 13-15 为描述统计量表。从表中可以看到，体能测试的平均成绩为 71.88。

表 13-15 描述统计量表

Descriptive Statistics								
	N	Mean	Std.Deviation	Minimum	Maximum	Percentiles		
						25th	50th (Median)	75th
体能测试成绩	16	71.88	15.811	38	103	62.75	72.00	81.50

表 13-16 为二项分布概率检验结果表，它给出了二项分布检验的最后结果。从表中可知，第一组“≤60”的观测量总数为 3，比例为 0.2，检验比例为 0.1，单尾显著性概率为 0.211，显然大于 0.05。

表 13-16 二项分布概率检验结果表

Binomial Test						
	Category	N	Observed Prop.	Test Prop.	Exact Sig. (1-tailed)	
体能测试成绩	Group 1 ≤60	3	0.2	0.1	0.211	
	Group 2 >60	13	0.8			
	Total	16	1.0			

因此得出结论：不拒绝零假设，即该批学生体能测试及格率可能达到 90% 了。

13.2.3 游程检验——Runs过程

游程检验是利用游程的总个数获得统计推断结论的方法。先引入以下概念：在一个二元序列中，一个由 0 和 1 连续构成的串称为一个游程，一个游程中数据的个数称为游程的长度。比如序列如下：

1110000111100100000

在这个序列中，111、0000、1111、00、1、00000 都是游程，其中第一个游程 111 的长度为 3。假设用 U 表示序列中游程的总数，用 V 表示最大游程长度。游程检验就是借助于 U 值和 V 值而建立起来的，用于检验两个总体是否相同，以及检验一个样本随机性的非参数检验法。

1. 游程检验的数据要求及界面操作简介

游程检验过程要求检验变量必须是数量型的。检验不需要有关分布类型的假设，可以

使用连续型分布的样本。

执行【Analyze】/【Nonparametric Tests】/【Runs】命令，弹出如图 13-7 所示的【Runs Test】（游程检验）对话框。下面来介绍其中主要元素。

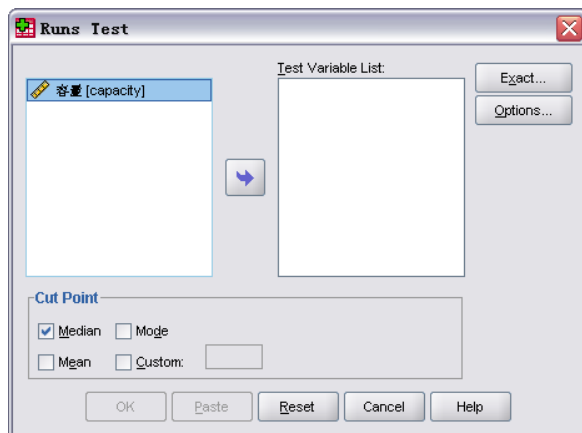


图 13-7 【游程检验】对话框

- 【Test Variable List】框

检验变量列表框，放置检验变量。可以选入多个检验变量到其中，但是系统进行处理时，是对每一个变量分别进行处理。

- 【Cut Point】复选框

断点设置复选框。系统提供了 4 种不同的断点设置方式，分别为 Median（中位数）、Mode（众数）、Mean（均值）和 Custom（用户自定义断点值），在其后的空白栏内输入这个断点值。对每一个选中的断点，系统都将分别进行检验。

- 【Exact Tests】子对话框

单击【Exact】按钮，弹出【Exact Tests】子对话框。

- 【Options】子对话框

单击【Options】按钮，弹出【Options】子对话框。

上面两个子对话框与图 13-3 和图 13-4 所示的子对话框一致。

2. 引例及结果解释

例 13.5 样本随机性检验。某品牌消毒液质检部要求每瓶消毒液的平均容积为 500ml，现从流水线上的某台装瓶机上随机抽取 20 瓶，测得其容量如表 13-17 所示，试检查这台机器装多装少是否随机？

表 13-17 某品牌消毒液每瓶容量抽查结果

509	505	502	501	493	498	497	502	504	506
505	508	498	495	496	507	506	507	508	505

假设检验问题：

H_0 : 机器装多装少是随机的。

STEP 01 建立数据文件“某品牌消毒液每瓶容量抽查.sav”，建立变量“capacity”——容量，将以上数据依次输入。

STEP 02 利用游程检验检验机器装多装少是否随机。

执行【Analyze】/【Nonparametric Tests】/【Runs】命令，弹出【Runs Test】对话框

Test Variable List : capacity	选入检验变量
Cut Point : Median	选择中位数作为断点
单击【Ok】按钮	生成以下结果

表 13-18 为游程检验结果表，它显示了游程检验的最后结果。从表中可以看出，游程检验的检验值为 505，本例里即是观测值的中位数为 505，观测量总数为 20，其中小于检验值的观测量总数为 10，其余的观测量总数也为 10，游程总数为 5，Z 检验统计量值为 -2.527，渐进的双尾显著性概率为 0.012，小于 0.05。

表 13-18 游程检验结果表

Runs Test	
	容 量
Test Value ^a	505
Cases < Test Value	10
Cases ≥ Test Value	10
Total Cases	20
Number of Runs	5
Z	-2.527
Asymp. Sig. (2-tailed)	0.012

a. Median

因此得出结论：拒绝零假设，认为这台机器装多装少并非随机的。

例 13.6 比较两个总体的检验。假定有 5 位健康成年人的血液，测量血液中的尿酸浓度，分别用手工和仪器两种方式测量，结果如表 13-19 所示，问这两种测量方法的精确度是否存在差异。

表 13-19 血液中尿酸浓度测量

手工 (X)	4.5	6.5	7	10	12
仪器 (Y)	6	7.2	8	9	9.8

假设检验：

H_0 ：总体 X 和总体 Y 具有相同的分布，即两种测量方法没有明显的差异。

STEP 01 建立数据文件“血液中尿酸浓度测量.sav”，将两组数据混合，建立两个变量，“blood”——血液中尿酸浓度和“group”——不同测量方法。变量“group”中用数值 1 代表手工，用数值 2 代表仪器。

STEP 02 将数据排序。

执行【Data】/【Sort Cases】命令，弹出【Sort Cases】对话框

Sort by : blood	对数据排序
单击【OK】按钮	观测量将按照变量“ blood ”的升序排列

STEP 03 利用游程检验检验两个分组是否具有相同的分布。

执行【Analyze】/【Nonparametric Tests】/【Runs】命令，弹出【Runs Test】对话框	
Test Variable List : group	选入检验变量
Cut Point : 勾选 Mean	选择均值作为断点
单击【OK】按钮	生成以下结果

表 13-20 为游程检验结果表，它显示了游程检验的最后结果。从表中可以看出，游程检验的检验值为 1.50，本例里即是观测值的均值为 1.50，观测量总数为 10，其中小于检验值的观测量总数为 5，其余的观测量总数也为 5，游程总数为 5，Z 检验统计量值为-0.335，渐进的双尾显著性概率为 0.737，远大于 0.05。

表 13-20 游程检验结果表

Runs Test	
	不同测量方法
Test Value ^a	1.50
Cases < Test Value	5
Cases ≥ Test Value	5
Total Cases	10
Number of Runs	5
Z	-0.335
Asymp. Sig. (2-tailed)	0.737

a. Mean

因此得出结论：不拒绝零假设，两种测量方法没有明显的差异。

13.2.4 单个样本的K-S检验——1-Sample K-S过程

K-S 检验就是 Kolmogorov-Smirnov 检验的简称，它的检验方法是以样本数据的累计频数分布与某个特定的理论分布相比较，若两者间的差距很小，则推论该样本取自某特定分布族。

对于连续型变量，在卡方拟合优度检验中需要人为地划分样本空间为区间或者分类，这样可能因为区间划分的不同而导致对同一个样本的检验得出完全对立的检验结果，而 K-S 检验在一定程度上克服了这个问题，因此，它比卡方检验更加精确。

1. 单个样本的K-S检验的统计原理

这里以 K-S 正态性检验为例介绍它的统计原理。

假设检验问题：

H_0 ：样本所来自的总体分布服从正态分布；

H_1 ：样本所来自的总体分布不服从正态分布。

$F_0(x)$ 表示分布的分布函数, $F_n(x)$ 表示一组随机样本的累计概率函数。设 D 为 $F_0(x)$ 与 $F_n(x)$ 差距的最大值, 定义如下公式:

$$D = \max |F_n(x) - F_0(x)|$$

结论: 当实际观测 $D > D_\alpha$ (其中 α 为给定的显著性水平) 时, 则接受 H_1 , 反之则不拒绝 H_0 假设。

2. 单个样本的K-S检验的数据要求及界面操作简介

K-S 检验过程要求检验变量为区间或者比例测度为数值型变量。

执行【Analyze】/【Nonparametric Tests】/【1-Sample K-S】命令, 弹出如图 13-8 所示的【One-Sample Kolmogorov-Smirnov Test】(单个样本的 K-S 检验) 对话框。下面介绍其中主要元素。

- 【Test Variable List】框

检验变量列表框, 放置检验变量。可以选入多个检验变量到其中, 但是系统进行处理时, 是对每一个变量分别进行处理。

- 【Test Distribution】复选框

检验的概率分布复选框, 从中选择需要检验的概率分布。系统提供了 4 种常见的分布, 分别是 Normal (正态分布)、Uniform (均匀分布)、Poisson (泊松分布) 和 Exponential (指数分布), 系统默认选项为 Normal。对每一个被选中的分布, 系统都将分别进行检验。

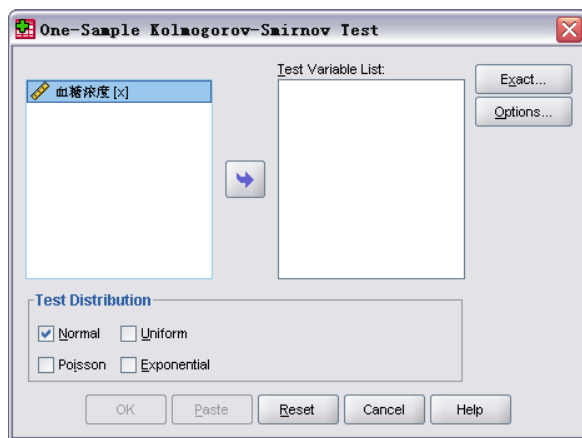


图 13-8 【单个样本的 K-S 检验】对话框

- 【Exact Tests】子对话框

单击【Exact】按钮, 弹出【Exact Tests】子对话框。

- 【Options】子对话框

单击【Options】按钮, 弹出【Options】子对话框。

上面两个子对话框与图 13-3 和图 13-4 所示的子对话框一致。

3. 引例及结果解释

例 13.7 K-S 正态性检验。35 位健康成年男性在未进食前的血糖浓度如表 13-21 所示，试检测这组数据是否服从正态分布。

表 13-21 血糖浓度抽查

87	77	92	68	80	78	84	77	81	80	80	77	92	86
76	80	81	75	77	72	81	90	84	86	80	68	77	87
76	77	78	92	75	80	78							

假设检验问题：

H_0 ：健康成年男性血糖浓度服从正态分布。

H_1 ：健康成年男性血糖浓度不服从正态分布。

STEP 01 建立数据文件“血糖浓度抽查.sav”。建立变量“x”——血糖浓度，将以上数据依次输入。

STEP 02 利用单样本的 K-S 检验文件中数据是否服从正态分布。

执行【Analyze】/【Nonparametric Tests】/【1-Sample K-S】命令，弹出对话框

Test Variable List : x

选入检验变量

Test Distribution : Normal

选择正态分布检验

单击【OK】按钮

生成以下结果

表 13-22 为单样本 K-S 检验结果表，它给出了单样本 K-S 检验的最后结果。从表中可以看出，总观测测量数为 35，正态分布参数中均值为 80.26，标准差为 6.026，最大极差的绝对值为 0.165，最大正极端差为 0.165，最大负极端差为 -0.106，K-S 检验统计量 Z 值为 0.978，双尾渐进显著性概率为 0.295，大于 0.05。

表 13-22 单样本 K-S 检验结果表

One-Sample Kolmogorov-Smirnov Test

		血糖浓度
N		35
Normal Parameters ^{a,b}	Mean	80.26
	Std. Deviation	6.026
Most Extreme Differences	Absolute	0.165
	Positive	0.165
	Negative	-0.106
Kolmogorov-Smirnov Z		0.978
Asymp. Sig. (2-tailed)		0.295

a. Test distribution is Normal.

b. Calculated from data.

因此得出结论：不拒绝零假设，表示健康成年男性血糖浓度可能服从正态分布。

13.3 分布位置检验

检验样本所在总体的分布位置或者形状是否相同。SPSS 为这一类检验也提供了 4 个子菜单，本节将通过例子详细介绍这 4 种检验方法的一般步骤、操作界面及结果解释。

13.3.1 两个独立样本分布位置检验——2 Independent Samples过程

两个独立样本分布位置检验用于当样本所属总体的分布类型不明时，检验两个独立样本是否来自相同的分布总体。

1. 两个独立样本分布位置检验的界面操作简介

执行【Analyze】/【Nonparametric Tests】/【2 Independent Samples】命令，弹出如图 13-9 所示的【Two-Independent-Samples Tests】（两个独立样本分布位置检验）对话框。下面介绍其中主要元素。

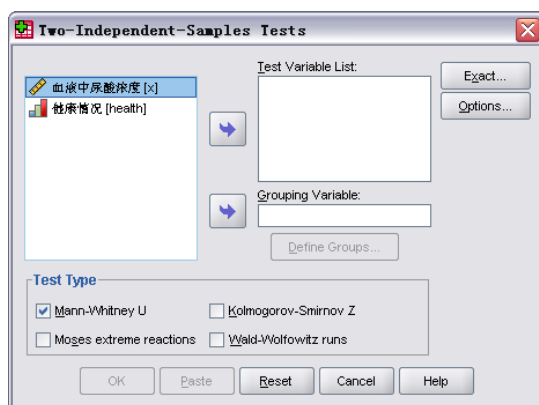


图 13-9 【两个独立样本分布位置检验】对话框

- 【Test Variable List】框：检验变量列表框，放置检验变量。
 - 【Grouping Variable】框：分组变量框，放置分组变量。选入分组变量后，单击【Define Range】按钮，弹出【Define Groups】子对话框，在其中输入两个分组的取值。在这两个取值外的观测值将被排除在检验分析之外。
 - 【Test Type】复选框：选择检验方法。具体方法将在下面介绍。
 - 【Exact Tests】子对话框：单击【Exact】按钮，弹出【Exact Tests】子对话框。
 - 【Options】子对话框：单击【Options】按钮，弹出【Options】子对话框。
- 上面两个子对话框与图 13-3 和图 13-4 所示的子对话框一致。

2. 检验方法基本原理

SPSS 为两个独立样本分布位置检验提供了 4 种检验方法。

- Mann-Whitney U: Mann-Whitney 的 U 检验法, 即 Wilcoxon 秩和检验法。该检验法是最常用的一种方法, 是系统默认检验方法。这种方法要求数据为 Ordinal 测度水平。
- Moses extreme reactions: Moses 极端反映检验法。如果施加的处理使得某些个体出现正向效应, 而另一些个体出现负向效应时, 就应当采用该检验方法。
- Kolmogorov-Smirnov Z: K-S 的 Z 检验法。
- Wald-Wolfowitz runs: Wald-Wolfowitz 游程检验法。它需要数据类型为 Ordinal 测度水平。

3. 引例及结果解释

例 13.8 两个独立样本检验实例。中风病人与正常人血液中尿酸浓度如表 13-23 所示。研究中风病人与正常人血液中尿酸浓度是否有明显的差异。

表 13-23 中风病人和正常人血液中尿酸浓度调查表

病人	8.2	10.7	7.5	14.6	6.3	9.2	11.9	5.6	12.8	5.2	4.9	13.5
正常人	4.7	6.3	5.2	6.8	5.6	4.2	6.0	7.4	8.1	6.5		

假设检验:

H_0 : 中风病人与正常人血液中尿酸浓度没有明显的差异。

STEP 01 建立数据文件“病人与正常人血液中尿酸浓度表.sav”, 将两组数据混合, 建立两个变量, “x”——血液中尿酸浓度和“health”——健康情况。变量“health”中用数值 1 代表病人, 数值 2 代表正常人。

STEP 02 将数据排序。

执行【Data】/【Sort Cases】命令, 弹出【Sort Cases】对话框

Sort by : x	对数据排序
单击【OK】按钮	观测量将按照变量“x”的升序排列

STEP 03 采用 Mann-Whitney U 检验和 K-S 的 Z 检验方法。

执行【Analyze】/【Nonparametric Tests】/【2 Independent Samples】命令, 弹出对话框

Test Variable List : x	选入检验变量
Grouping Variable : health (1 , 2)	选入分组变量, 并确定好分组的取值
Test Type : Mann-Whitney U	采用 Mann-Whitney U 检验法
勾选 Kolmogorov-Smirnov Z	采用 K-S 的 Z 检验法
单击【Options】按钮	进入【Options】子对话框
Statistics : 勾选 Descriptive	输出描述统计量
勾选 Quartiles	输出四分位数
单击【Continue】按钮	回到主对话框
单击【Ok】按钮	生成以下结果

• Npar Tests 部分。表 13-24 为描述统计量表。

表 13-24 描述统计量表

Descriptive Statistics								
	N	Mean	Std.Deviation	Minimum	Maximum	Percentiles		
						25th	50th (Median)	75th
血液中尿酸浓度	22	7.782	3.0551	4.2	14.6	5.500	6.650	9.575
健康情况	22	1.45	0.510	1	2	1.00	1.00	2.00

- Mann-Whitney Test 部分，即 Mann-Whitney 检验部分，包括表 13-25 和表 13-26。表 13-25 中列出了两个样本的秩和及平均秩。

表 13-25 Mann-Whitney 检验的秩统计量表

Ranks				
健康情况	N	Mean Rank	Sum of Ranks	
病人	12	14.21	170.50	
正常人	10	8.25	82.50	
Total	22			

表 13-26 为 Mann-Whitney 检验的最后结果表，从表中可见 Mann-Whitney 的 U 统计量值等于 27.000，Wilcoxon 的 W 统计量值为 82.500，Z 值等于-2.145，双尾渐进显著性概率为 0.032，小于 0.05。

表 13-26 Mann-Whitney 检验结果表

Test Statistics ^b	
	血液中尿酸浓度
Mann-Whitney U	27.500
Wilcoxon W	82.500
Z	-2.145
Asymp. Sig. (2-tailed)	0.032
Exact Sig. [2*(1-tailed Sig.)]	0.030 ^a

a. Not corrected for ties.

b. Grouping Variable: 健康情况

因此得出结论：拒绝零假设，病人和正常人血液里的尿酸浓度有明显的差异。

- Two-Sample K-S Test 部分。两个样本的 K-S 检验部分，包括表 13-27 和表 13-28 两个表。

表 13-27 K-S Z 检验频数表

Frequencies		
健康情况	N	
血液中尿酸浓度	病人	12
	正常人	10
	Total	22

表 13-28 K-S Z 检验结果表

Test Statistics ^a		血液中尿酸浓度
Most Extreme Differences	Absolute	0.583
	Positive	0.000
	Negative	-0.583
Kolmogorov-Smirnov Z		1.362
Asymp. Sig. (2-tailed)		0.049

a. Grouping Variable: 健康情况

从表 13-28 中可以看到 K-S Z 检验的双尾显著性概率为 0.049，小于 0.05。因此得出相同的结论：拒绝零假设，病人和正常人血液里的尿酸浓度有明显的差异。

13.3.2 多个独立样本分布位置检验——K Independent Samples 过程

多个独立样本分布位置检验是要解决多个独立样本间是否具有相同分布的问题。

1. 多个独立样本分布位置检验的基本步骤及界面操作简介

执行【Analyze】/【Nonparametric Tests】/【K Independent Samples】命令，弹出如图 13-10 所示的【Tests for Several Independent Samples】（多个独立样本分布位置检验）对话框。下面介绍其中的主要元素。

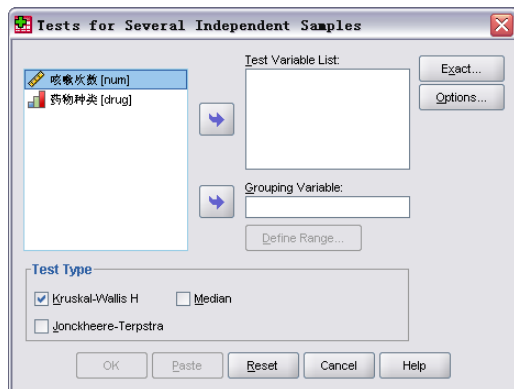


图 13-10 【多个独立样本分布位置检验】对话框

- 【Test Variable List】框：检验变量列表框，放置检验变量。
- 【Grouping Variable】框：分组变量框，放置分组变量。选入分组变量后，单击【Define Range】按钮，弹出【Define Groups】子对话框，在其中输入分组的取值范围。在这个取值范围外的观测值将被排除在检验分析之外。
- 【Test Type】复选框：选择检验方法。
- 【Exact Tests】子对话框：单击【Exact】按钮，弹出【Exact Tests】子对话框。

- **【Options】子对话框：**单击**【Options】**按钮，弹出**【Options】**子对话框。

上面两个子对话框与图 13-3 和图 13-4 所示的子对话框一致。

2. 检验方法基本原理

SPSS 为多个独立样本分布位置检验提供了三种检验方法。

- **Kruskal-Wallis H：**Kruskal-Wallis H 检验法，它是 Mann-Whitney U 检验法的推广，是类似于单因素方差分析的一种检验法。系统默认检验法，要求总体是连续性随机变量，至少具有 Ordinal 测度水平。
- **Median：**中位数检验法。该方法是使用非常普遍的一种检验法。
- **Jonckheere-Terpstra：**Jonckheere-Terpstra 检验法，用于解决位置参数某一方向的假设检验问题。

3. 引例及结果解释

例 13.9 多个独立样本检验实例。为研究 4 种不同药物对儿童咳嗽的治疗效果，将 25 个体质相似的病人随机分为 4 组，各组人数分别为 8 人、4 人、7 人和 6 人，各自采用 A、B、C、D 四种药物进行治疗，假定其他条件均保持相同，5 天后测量每个病人每天的咳嗽次数如表 13-29 所示，试比较这四种药物的治疗效果是否相同。

表 13-29 四种药物治疗效果比较表

	A	B	C	D
1	80	133	156	194
2	203	180	295	214
3	236	100	320	272
4	252	160	448	330
5	284		465	386
6	368		481	475
7	457		279	
8	393			

假设检验：

H_0 ：四种药物的治疗效果相同。

STEP 01 建立数据文件“四种药物治疗效果比较表.sav”，建立两个变量，“num”——咳嗽次数和“drug”——药物种类。变量“drug”中用数值 1 代表药物 A，数值 2 代表药物 B，数值 3 代表药物 C，数值 4 代表药物 D。

STEP 02 采用 Kruskal-Wallis H 检验法和中位数检验法。

执行**【Analyze】** **【Nonparametric Tests】** **【K Independent Samples】**命令，弹出**【K Independent Samples】**对话框

Test Variable List：num

选入检验变量

Grouping Variable：drug（1，4）

选入分组变量，并确定好分组的取值范围

Test Type：Kruskal-Wallis H

采用 Kruskal-Wallis H 检验法

勾选 Median

采用中位数检验法

单击**【OK】**按钮

生成以下结果

- Kruskal-Wallis Test 部分。Kruskal-Wallis H 检验部分，包括表 13-30 和表 13-31。

表 13-30 平均秩表

Ranks			
药物种类		N	Mean Rank
咳嗽次数	A	8	13.00
	B	4	4.00
	C	7	16.71
	D	6	14.67
	Total	25	

表 13-31 H 检验结果表

Test Statistics ^{a,b}	
	咳嗽次数
Chi-Square	8.072
Df	3
Asymp. Sig.	0.045

a. Kruskal Wallis Test

b. Grouping Variable: 药物种类

由表 13-31 可以得出 H 检验结论：拒绝零假设，4 种药物的治疗效果有差异。

- Median Test 部分。中位数检验部分，包括表 13-32 和表 13-33。

表 13-32 频数表

Frequencies					
		药物种类			
		A	B	C	D
咳嗽次数	> Median	4	0	5	3
	□ Median	4	4	2	3

表 13-33 中位数检验结果表

Test Statistics ^b	
	咳嗽次数
N	25
Median	279.00
Chi-Square	5.254 ^a
df	3
Asymp. Sig.	0.154

a. 8 cells (100.0%) have expected frequencies less than 5. The minimum expected cell frequency is 1.9.

b. Grouping Variable: 药物种类

由表 13-33 得出中位数检验结论：接受零假设，4 种药物的治疗效果相同。

结合两种方法进行分析，因为 H 检验时，渐进显著性概率值为 0.045 很接近 0.5，所以，基本上认为这 4 种药物疗效区别不大。

13.3.3 两个相关样本分布位置检验——2 Relate Samples过程

当两个样本间的数据不再是相互独立，而是彼此相关时，可以用两个相关样本分布位置检验来检验两个样本是否有相同的分布。

1. 两个相关样本分布位置检验的基本步骤及界面操作简介

执行【Analyze】/【Nonparametric Tests】/【2 Relate Samples】命令，弹出如图 13-11 所示的【Two-Related-Samples Tests】（两个相关样本分布位置检验）对话框。下面介绍其中的主要元素。

- 【Test Pairs List】框：配对检验变量列表框，放置配好对的检验变量。
- 【Current Selections】框：当前选择框，显示当前选择的配对变量。选中原变量列表框中的一个变量后，此变量自动显示在 Variable 1 后，再选中另一个变量，此变量自动显示在 Variable 2 后，此时，单击箭头按钮，这两个变量配对出现在【Test Pairs List】框中。
- 【Test Type】复选框：选择检验方法。
- 【Exact Tests】子对话框：单击【Exact】按钮，弹出【Exact Tests】子对话框。
- 【Options】子对话框：单击【Options】按钮，弹出【Options】子对话框。

上面两个子对话框与图 13-3 和图 13-4 所示的子对话框一致。

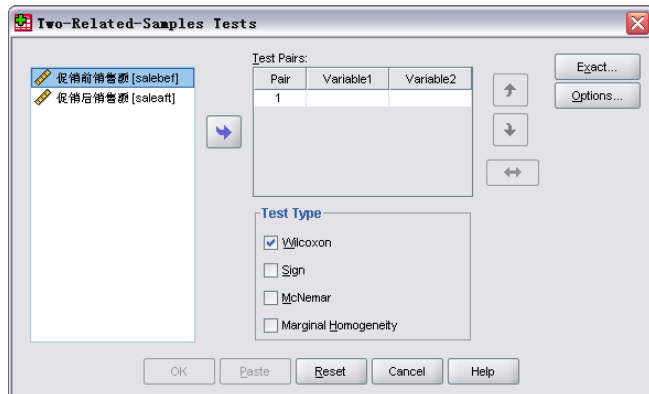


图 13-11 【两个相关样本分布位置检验】对话框

2. 检验方法基本原理

SPSS 为两个相关样本分布位置检验提供了 4 种检验方法。

- **Wicoxon:** Wicoxon 符号秩检验法，当变量为连续型变量时，选用此法。它是系统默认检验法。
- **Sign:** 符号检验法，这种方法也用于连续变量，但效果没有前者好。
- **McNemar:** McNemar 检验法，适合于两个相关的二元变量总体的检验。
- **Marginal Homogeneity:** 边际同质检验法。是 McNemar 检验法的推广，适合于分类变量，而不仅仅限制在二元分类变量。

3. 引例及结果解释

例 13.10 两个相关样本检验实例。表 13-34 是某种商品在 12 家超市促销活动前后的销售额比较数据，试检验分析促销活动的效果如何。

表 13-34 某种商品促销活动前后销售额的比较表

连锁店	促销前销售额	促销后销售额
1	42	40
2	57	60
3	38	38
4	49	47
5	63	65
6	36	39
7	48	49
8	58	50
9	47	47
10	51	52
11	83	72
12	27	33

假设检验：

H_0 : 促销前后的销售额无明显差异。

STEP 01 建立数据文件“促销活动前后销售额的比较表.sav”，建立两个变量，“salebef”——促销前销售额和“saleaft”——促销后销售额。

STEP 02 采用 Wicoxon 符号秩检验法和符号检验法。

执行【Analyze】/【Nonparametric Tests】/【2 Relate Samples】命令，弹出【Two-Related-Samples Tests】对话框

Test Pairs List : salebef - saleaf

选入配对检验变量

Test Type : Wicoxon

采用 Wicoxon 符号秩检验法

勾选 Sign

采用符号检验法

单击【Options】按钮

进入【Options】子对话框

Statistics : 勾选 Descriptive

输出描述统计量

单击【Continue】按钮

回到主对话框

单击【Ok】按钮

生成以下结果

- Npar Tests 部分。表 13-35 为描述统计量表。

表 13-35 描述统计量表

Descriptive Statistics					
	N	Mean	Std. Deviation	Minimum	Maximum
促销前销售额	12	49.92	14.519	27	83
促销后销售额	12	49.33	11.602	33	72

- Wilcoxon Signed Ranks Test 部分。Wilcoxon 符号秩检验法部分，由表 13-36 和表 13-37 构成。

表 13-36 秩计算结果表

Ranks		N	Mean Rank	Sum of Ranks
促销后销售额 -	Negative Ranks	4 ^a	6.75	27.00
促销前销售额	Positive Ranks	6 ^b	4.67	28.00
	Ties	2 ^c		
	Total	12		

a. 促销后销售额 < 促销前销售额

b. 促销后销售额 > 促销前销售额

c. 促销后销售额 = 促销前销售额

表 13-37 Wilcoxon 符号秩检验法结果表

Test Statistics ^b	
	促销后销售额-促销前销售额
Z	-0.051 ^a
Asymp. Sig. (2-tailed)	0.959

a. Based on negative ranks.

b. Wilcoxon Signed Ranks Test

由表 13-37 可以看出，双尾渐进显著性概率值为 0.959，远远大于 0.05，得出结论：不能拒绝零假设，没有证据显示促销前后的销售额有明显差异。

- ③ Sign Test 部分。符号检验部分。由表 13-38 和表 13-39 构成。

表 13-38 符号检验的频数表

Frequencies		N
促销后销售额 -	Negative Differences ^a	4
促销前销售额	Positive Differences ^b	6
	Ties ^c	2
	Total	12

a. 促销后销售额 < 促销前销售额

b. 促销后销售额 > 促销前销售额

c. 促销后销售额 = 促销前销售额

表 13-39 符号检验结果表

Test Statistics ^b	
	促销后销售额-促销前销售额
Exact Sig. (2-tailed)	0.754 ^a

a. Binomial distribution used.

b. Sign Test

从表 13-39 可以得其结论：不能拒绝零假设，没有证据显示促销前后的销售额有明显差异，与 Wilcoxon 符号秩检验法结论一致。

13.3.4 多个相关样本分布位置检验——K Relate Samples过程

顾名思义，多个相关样本分布位置检验就是用来检验多个相关样本间是否有相同的分布。

1. 多个相关样本分布位置检验的基本步骤及界面操作简介

执行【Analyze】/【Nonparametric Tests】/【K Relate Samples】命令，弹出如图 13-12 所示的【Tests for Several Related Samples】（多个相关样本分布位置检验）对话框。下面介绍其中主要元素。

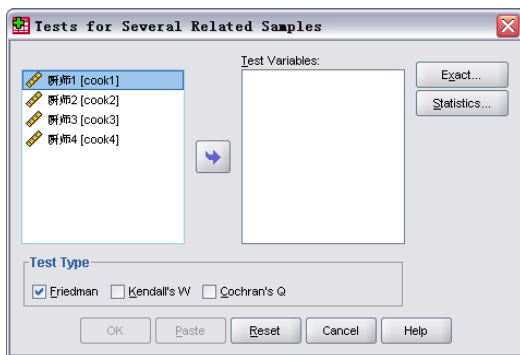


图 13-12 【多个相关样本分布位置检验】对话框

- 【Test Variable】框：检验变量列表框，放置检验变量。
- 【Test Type】复选框：选择检验方法。
- 【Exact Tests】子对话框：单击【Exact】按钮，弹出【Exact Tests】子对话框。此子对话框与图 13-3 所示的子对话框一致。
- 【Statistics】子对话框：单击【Statistics】按钮，弹出如图 13-13 所示的【Statistics】子对话框。该对话框主要用来定义输出的统计量信息。

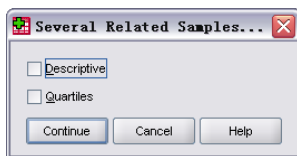


图 13-13 【Statistics】子对话框

2. 检验方法基本原理

SPSS 为多个相关样本分布位置检验提供了三种检验方法。

- **Friedman:** Friedman 检验法，对每一变量的观测值，赋予 1-k 的秩，基于这些秩确定检验的 Friedman 统计量。它是系统默认检验法。
- **Kendall's W:** Kendall 检验法，其中的 W 统计量是对 Friedman 统计量的正态化。
- **Cochran's Q:** Cochran 检验法，是一种检验二元变量总体均值是否相等的非参数检验方法。

3. 引例及结果解释

例 13.11 多个相关样本检验实例。设有来自 A、B、C、D 四个地区的四名厨师制作名菜水煮鱼，想比较它们的品质是否相同，四位美食评委评分结果见表 13-40，试测试四个地区制作的水煮鱼这道菜品品质有无区别。

表 13-40 评委对四名厨师的评分数据表

美食评委	地 区			
	A	B	C	D
1	85	82	82	79
2	87	75	86	82
3	90	81	80	76
4	80	75	81	75

假设检验：

H_0 ：四个地区的水煮鱼品质没有区别。

STEP 01 建立数据文件“评委对四名厨师的评分数据表.sav”，建立四个变量，“cook1”——厨师 1 得分、“cook2”——厨师 2 得分、“cook3”——厨师 3 得分和“cook4”——厨师 4 得分。

STEP 02 采用 Friedman 检验法和 Kendall 检验法。

执行【Analyze】/【Nonparametric Tests】/【K Relate Samples】命令，弹出图 13-12 所示对话框

Test Variables : cook1、cook2、cook3、cook4

选入检验变量

Test Type : Friedman

采用 Friedman 检验法

勾选 Kendall 's W

采用 Kendall 检验法

单击【Options】按钮

进入【Options】子对话框

Statistics : 勾选 Descriptive

输出描述统计量

单击【Continue】按钮

回到主对话框

单击【Ok】按钮

生成以下结果

- Npar Tests 部分。表 13-41 为描述统计量表。

表 13-41 描述统计量表

Descriptive Statistics					
	N	Mean	Std. Deviation	Minimum	Maximum
厨师 1	4	85.50	4.203	80	90
厨师 2	4	78.25	3.775	75	82
厨师 3	4	82.25	2.630	80	86
厨师 4	4	78.00	3.160	75	82

- Friedman Test 部分。Friedman 检验部分，包括表 13-42 和表 13-43。

表 13-42 平均秩表

Ranks	
	Mean Rank
厨师 1	3.75
厨师 2	2.00
厨师 3	2.88
厨师 4	1.38

表 13-43 Friedman 检验结果表

Test Statistics ^a	
N	4
Chi-Square	8.132
df	3
Asymp. Sig.	0.043

a. Friedman Test

从表 13-43 得出结论：拒绝零假设，四个地区的厨师制作的水煮鱼这道菜在品质上存在明显的差异。

- Kendall's W Test 部分。Kendall's W 检验部分，包括表 13-44 和表 13-45 表格。

表 13-44 平均秩表

Ranks	
	Mean Rank
厨师 1	3.75
厨师 2	2.00
厨师 3	2.88
厨师 4	1.38

表 13-45 Kendall's W 检验结果表

Test Statistics	
N	4
Kendall's W ^a	0.678
Chi-Square	8.132
df	3
Asymp. Sig.	0.043

a. Kendall's Coefficient of Concordance

从表 13-45 得出结论：拒绝零假设，四个地区的厨师制作的水煮鱼这道菜在品质上存在明显的差异。

13.4 本章小结

本章介绍了非参数检验在 SPSS 中的实现，详细介绍了分布类型和分布位置的检验，主要包括以下内容：

- Chi-Square 过程，卡方检验；
- Binomial 过程，二项分布检验；
- Runs 过程，游程检验；
- 1-Sample K-S 过程，单个样本的 K-S 检验；
- 2 Independent Samples 过程，两个独立样本分布位置检验；
- K Independent Samples 过程，多个独立样本分布位置检验；
- 2 Relate Samples 过程，两个相关样本分布位置检验；
- K Relate Samples 过程，多个相关样本分布位置检验。

在具体实现的时候，需要注意两点，第一是知道这些方法各自的适用条件；第二是要明确要处理的问题的零假设到底是什么。这两点是本书反复强调的地方，也是在统计分析中必须引起高度重视的地方。

PART

第 4 篇 应用实例

- 第 14 章 SPSS 在各领域的应用实例

第 14 章 SPSS在各领域的应用实例

作为全球应用最广泛的统计软件，SPSS 已有 30 余年的历史。它广泛应用于农业、工业、商业、医学、交通运输、公检法、社会学、市场分析、股市行情、军事地理和旅游业等多个行业和领域。

在本书的开篇就曾提过，凡是需要数据分析的地方，都可以用到 SPSS。那么，在数据分析过程中，SPSS 究竟发挥着怎样的作用呢？SPSS 好比一件不错的工具，提供了多种成熟的分析方法。但是其作用的发挥还依赖于使用者的智慧。真正实际的统计问题都是没有一般性的方法和结论的。这既是统计学和传统数学的区别，也是统计学的魅力所在。成功解决一个问题，依赖于分析人员的智慧、对实际问题的理解以及对软件地分析操作。以上三者缺一不可。因此，可以这么说，在数据分析的过程中，SPSS 不是万能的。但是没有 SPSS 或其他统计软件却是万万不能的。

通过前面的学习，相信读者已经掌握了 SPSS 常用统计功能的基本操作。在本章，将通过多个具体例子，展示如何通过 SPSS 来处理实际问题。这些问题既有编者做过的具体模型，也有其他统计工作者的成功案例。希望这些案例的展示能对读者的实际工作有所启发。

本章精选了 SPSS 典型应用领域的实际例子，提供全面的问题解决方案，深入到 SPSS 实际的应用。

- SPSS 在房地产决策中的应用
- SPSS 在生物模型中的应用
- SPSS 在工程问题中的应用
- SPSS 在证券分析中的应用

14.1 SPSS在房地产决策中的应用

任何一个成熟的行业，其决策都依赖于统计分析。这是因为数据是真实客观的。但是在我国的房地产行业中，统计分析还处于一个初级阶段，有待进一步发展。其实在房地产业中，大到国家宏观调控，小到楼盘价格、楼盘定位等，统计分析都大有用武之地。本节给出 SPSS 在房地产业宏观决策中的一个应用实例。

14.1.1 问题描述

近年来,一些地区房地产投资增幅偏快、房价上涨过高。在房价上涨比较快的地方,伴随着商品房结构的不合理,中低价位、中小户型房屋严重供不应求。房价上涨与投机性购房相互推动:房价上涨带来了大家对房价不断上涨的期望值,投机和投资的人更多地进入,又进一步造成房地产的供不应求和房价的进一步攀升。这些问题带来了生产资料的价格上涨,涉及普通居民居住权益的保障,既影响国民经济的健康发展,也涉及社会的稳定。

由于目前各省市房地产业出现的普遍过热,宏观调控势在必行。然而各地区具体情况又参差不齐,宏观政策不能盲目地一概进行。因此,希望能够通过分析与各地区房地产业发展有关的经济、社会指标,对各省市的房地产业的供需水平进行初步地判断,从而制定相应的宏观调控政策。

14.1.2 问题建模

根据上面的问题,分别选取如表 14-1 和表 14-2 所示的 2002~2003 年的 12 个指标来分析问题。(数据来源:《中国统计年鉴 2002、2003、2004》,中国统计出版社)

表 14-1 2002~2003 年房地产业发展指标

省 市	02 年人均施工 面积 (平方米)	03 年人均施工 面积 (平方米)	02 年人均竣工 面积 (平方米)	03 年人均竣工 面积 (平方米)	02 年人均开发 投资 (元)	03 年人均开发 投资 (元)
北京	7.196697	8.22239	2.689389	3.025893	6952.916	8258.929
天津	2.798213	3.311771	1.32284	1.583581	1745.78	2090.999
河北	1.232056	1.291904	0.703964	0.617285	259.8367	372.2854
山西	0.884153	1.013971	0.435003	0.359475	204.6145	286.3609
内蒙	0.915637	1.022773	0.629298	0.528319	304.7499	381.9328
辽宁	1.721366	1.94658	0.960504	0.834822	923.8639	1154.869
吉林	0.917229	1.412204	0.591886	0.640496	432.7529	515.1627
黑龙江	0.872253	0.921756	0.556124	0.514181	382.3761	427.5229
上海	3.840062	4.995383	1.493415	1.637873	4608.615	5267.095
江苏	3.760114	4.731299	2.097087	2.311113	737.163	1092.628
浙江	6.135743	8.600427	2.750549	3.766645	1568.324	2077.778
安徽	1.000994	1.804602	0.607068	0.637629	231.1455	375.507
福建	1.598673	1.888389	0.72663	0.722964	718.4074	1038.131
江西	0.836286	1.443277	0.453742	0.629901	245.3813	408.5567
山东	1.671889	1.963912	0.888009	1.001359	430.7421	635.1781
河南	0.740497	0.81862	0.377697	0.324672	143.9717	191.9934
湖北	1.206997	1.171943	0.693253	0.675375	298.2632	398.2006
湖南	1.079484	1.34624	0.552285	0.586162	227.6361	345.1899
广东	2.402647	2.820493	1.011312	1.179092	1419.137	1521.121

续表

省 市	02 年人均施工 面积（平方米）	03 年人均施工 面积（平方米）	02 年人均竣工 面积（平方米）	03 年人均竣工 面积（平方米）	02 年人均开发 投资（元）	03 年人均开发 投资（元）
广西	0.621692	0.842928	0.274907	0.342557	156.1593	247.6838
海南	0.67858	0.509864	0.260025	0.066091	250.3113	442.6634
重庆	2.80251	2.877668	1.516286	1.28869	791.4387	1047.604
四川	1.718413	1.761529	0.979085	1.032046	397.0944	516.4368
贵州	0.601876	0.649819	0.274772	0.210904	216.3148	268.9922
云南	0.786084	0.805599	0.513778	0.410352	222.2479	256.3985
西藏	0.467041	0.25	0.355805	0.067037	104.8689	62.96296
陕西	0.905008	1.04981	0.395727	0.390244	336.4181	511.1111
甘肃	0.947435	0.977372	0.576822	0.453208	145.3914	195.1594
青海	0.853119	0.958614	0.461437	0.447191	319.4707	417.603
宁夏	1.459266	1.841552	0.753322	0.922414	540.2098	877.5862
新疆	1.350446	1.410858	0.771444	0.628438	460.3675	509.3071

表 14-2 居民人口数及储蓄情况

省 市	02 年总 人口（万）	03 年总 人口（万）	02 年年底 储蓄余额 （亿元）	03 年年底 储蓄余额 （亿元）	省 市	02 年总 人口（万）	03 年总 人口（万）	02 年年底 储蓄余额 （亿元）	03 年年底 储蓄余额 （亿元）
北京	1423	1456	4389.7	5293.5	湖北	5988	6002	2754.5	3296.5
天津	1007	1011	1486.4	1825.3	湖南	6629	6663	2576.4	3036.5
河北	6735	6769	4808.3	5457.0	广东	7859	7954	11813.3	14061.8
山西	3294	3314	2307.3	2781.5	广西	4822	4857	1733.5	1971.7
内蒙	2379	2380	1137.9	1355.5	海南	803	811	483.5	546.9
辽宁	4203	4210	4665.0	5434.7	重庆	3107	3130	1582.3	1896.6
吉林	2699	2704	1878.5	2161.4	四川	8673	8700	3665.2	4333.8
黑龙江	3813	3815	2915.7	3342.4	贵州	3837	3870	758.7	912.8
上海	1625	1711	3891.5	5103.2	云南	4333	4376	1499.8	1766.5
江苏	7381	7406	6276.2	7638.2	西藏	267	270	70.4	91.9
浙江	4647	4680	5212.7	6452.2	陕西	3674	3690	2108.1	2519.9
安徽	6338	6410	2047.5	2475.8	甘肃	2593	2603	1042.4	1217.4
福建	3466	3488	2430.5	2924.7	青海	529	534	222.4	260.5
江西	4222	4254	1706.6	2015.5	宁夏	572	580	306.8	377.7
山东	9082	9125	5803.5	6768.4	新疆	1905	1934	1137.6	1371.8
河南	9613	9667	4196.0	4919.1					

1. 创建SPSS数据文件并整理数据

将表 14-1 和表 14-2 的数据保存在数据文件“fangdichan.sav”中。该数据文件的变量名及其标签如图 14-1 所示。

Name	Label
SF	省份
RJSG02	02年人均施工面积
RJSG03	03年人均施工面积
RJJG02	02年人均竣工面积
RJJG03	03年人均竣工面积
RJTZ02	02年人均开发投资
RJTZ03	03年人均开发投资
ZRK02	02年总人口
ZRK03	03年总人口
CXZE02	02年储蓄总额
CXZE03	03年储蓄总额

图 14-1 数据文件“fangdichan.sav”的变量名及其标签

现在的问题是利用这些变量研究各省市房地产的供需情况。由于表 14-1 中的数据都是按人均计算的，因此有必要将表 14-2 中的储蓄量也按人均来计算。

这就需要引入新变量，执行以下操作：

执行【Transform】/【Compute Variable】命令，弹出【Compute Variable】对话框	
【Target Variable】框：RJCX02	定义生成新变量“RJCX02”
【Numeric Expression】框：CXZE02 / ZRK02	定义新变量的计算公式
单击【Type&Label】按钮	
【Type&Label】对话框：	
【Label】框：02 年人均储蓄额	定义新变量的变量标签
单击【Continue】按钮	变量标签定义完成
单击【OK】按钮	定义完成

执行以上操作之后，数据文件中将生成一系列新变量“RJCX02”，其变量标签为“02 年人均储蓄额”。重复以上操作，生成新变量“RJCX03”，其变量标签为“03 年人均储蓄额”。现将表 14-2 中的数据也转化到人均水平上，有利于考察问题。否则，在分析的时候会出现一些背离实际的结果。

2. 因子分析

本节的问题是研究房地产业的供需关系。但是目前包含的变量还是比较多的，不利于分析问题。因此一个自然的想法是用因子分析来研究能不能将多个变量综合为少数几个因子。执行以下操作：

执行【Analyze】/【Dimension Reduction】/【Factor】命令，弹出【Factor】对话框	
【Variable】框：RJSG02、RJJG02、RJTZ02、RJCX02、RJSG03、RJJG03、RJTZ03、RJCX03	定义因子分析的变量
单击【Descriptives】按钮	
【Descriptives】对话框：	
选中“KMO and Bartlett's test of sphericity”复选框	检验因子分析是否适用
默认选项不变，单击【Continue】按钮	该对话框定义完成
单击【Extraction】按钮	
【Extraction】对话框：	

选中“Scree plot”复选框	绘制碎石图
选中“Number of factors”单选框：2	定义提取因子数目
默认选项不变，单击【Continue】按钮	该对话框定义完成
单击【Rotation】按钮	弹出【Rotation】对话框
【Rotation】对话框：	
选中“Varimax”单选框	方差最大化旋转
选中“Loading plot”复选框	绘制因子载荷图
默认选项不变，单击【Continue】按钮	该对话框定义完成
单击【Scores】按钮	弹出【Scores】对话框
【Scores】对话框：	
选中“Save as variables”单选框	定义保存因子得分
默认选项不变，单击【Continue】按钮	该对话框定义完成
单击【OK】按钮	定义完成

执行以上操作之后，可以由生成的 KMO and Bartlett's 表（表 14-3）及 Communalities 表（表 14-4）判断数据能够进行因子分析，且选择的两个公共因子可以提取各变量 97% 以上的信息。具体的判别方法请参阅第 12 章的相应部分。

表 14-3 KMO and Bartlett's 检验结果表

KMO and Bartlett's Test		
Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		0.798
Bartlett's Test of Sphericity	Approx. Chi-Square	771.823
	df	28
	Sig.	0.000

表 14-4 公因子方差比表

Communalities		
	Initial	Extraction
02 年人均施工面积	1.000	0.994
03 年人均施工面积	1.000	0.988
02 年人均竣工面积	1.000	0.983
03 年人均竣工面积	1.000	0.993
02 年人均开发投资	1.000	0.979
03 年人均开发投资	1.000	0.975
02 年人均储蓄额	1.000	0.976
03 年人均储蓄额	1.000	0.976

Extraction Method: Principal Component Analysis.

表 14-5 是主成分列表。从表中可以看出第一主成分特征根为 6.906，解释了总变异的 86.327%；第二主成分特征根为 0.957，解释了总变异的 11.963%。因此对于原始的 8 个变量指标只需提取第一主成分和第二主成分即可。

表 14-5 主成分列表

Total Variance Explained

Component	Initial Eigenvalues			Extraction Sums of Squared Loadings			Rotation Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	6.906	86.327	86.327	6.906	86.327	86.327	4.009	50.108	50.108
2	0.957	11.963	98.291	0.957	11.963	98.291	3.855	48.182	98.291
3	0.096	1.200	99.491						
4	0.029	0.362	99.853						
5	0.007	0.083	99.936						
6	0.004	0.046	99.981						
7	0.001	0.012	99.994						
8	0.001	0.006	100.000						

Extraction Method: Principal Component Analysis.

当确定选择两个主成分之后，为了能够给出其合理的解释有必要进行一定的旋转。表 14-6 即为旋转后的因子负荷矩阵，图 14-2 为旋转后的主成分图。

表 14-6 旋转后的因子负荷矩阵

Rotated Component Matrix^a

	Component	
	1	2
02 年人均施工面积	0.530	0.844
03 年人均施工面积	0.471	0.875
02 年人均竣工面积	0.379	0.916
03 年人均竣工面积	0.331	0.940
02 年人均开发投资	0.905	0.399
03 年人均开发投资	0.890	0.428
02 年人均储蓄额	0.907	0.390
03 年人均储蓄额	0.905	0.397

Extraction Method: Principal Component Analysis.

Rotation Method: Varimax with Kaiser Normalization.

a. Rotation converged in 3 iterations.

Component Plot in Rotated Space

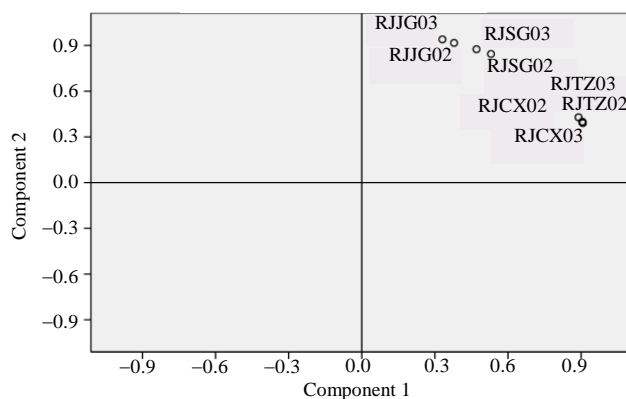


图 14-2 旋转后的主成分图

下面根据指标特点给出第一主成分和第二主成分的合理解释。从表 14-6 的因子负荷矩阵可以看出，所有变量均对第一、第二主成分有所贡献。但是对第一主成分贡献最大的是 02、03 年的人均开发投资和人均储蓄额指标。由于居民购房主要是基于居住和投资两大目的，而购房的先决条件是有了一定的储蓄，同时居民对房地产的需求在很大程度上又会加大投资商的投资力度。因此，第一主成分在很大程度上反映的是对房地产的需求程度。对第二主成分贡献最大的是 02、03 年的人均施工面积和竣工面积指标。可见，第二主成分反映的是房地产市场的一个供给水平。根据解释，可以将原来的 8 个变量按照其反映的问题分为如图 14-3 所示的两类。

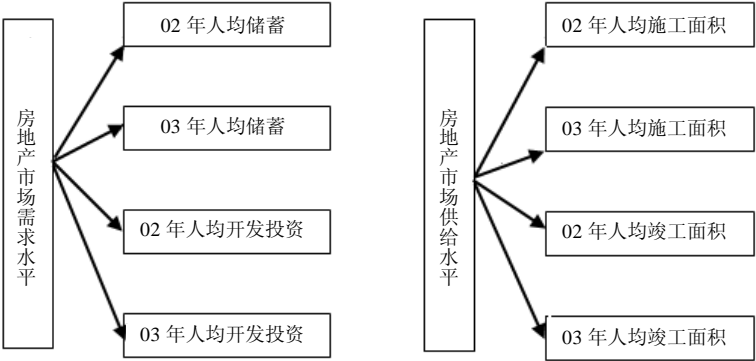


图 14-3 房地产市场供需指标分类

3. 根据因子得分情况进行聚类分析

进行如上所述的因子分析后，在原始的数据文件“fangdichan.sav”中会生成两列新变量“FAC1_1”和“FAC2_1”，用来保存各省份的因子得分情况。执行以下操作：

执行【Graph】/【Legacy Dialogs】/【Scatter】命令，选择【Simple Scatter】 定义绘制简单散点图	
【Simple Scatter】对话框：	
【Y Axis】框：FAC2_1	定义简单散点图 Y 轴
【X Axis】框：FAC1_1	定义简单散点图 X 轴
【Set Markers by】框：SF	定义标识变量
单击【OK】按钮	定义完成

执行以上操作并对图形进行简单编辑后，有如图 14-4 所示散点图。
从图 14-4 上看，可以将各省市的情况大致分成三类，执行以下操作：

执行【Analyze】/【Classify】/【K-means Cluster】命令，弹出【K-means Cluster】对话框	
【Variable】框：FAC1_1、FAC2_1	定义按照因子得分进行聚类
【Number of Clusters】框：3	定义聚类数
单击【Save】按钮	弹出【Save】对话框
选中“Cluster membership”复选框	定义保存各记录所属类别
单击【Continue】按钮	【Save】对话框定义完成

单击【OK】按钮

定义完成

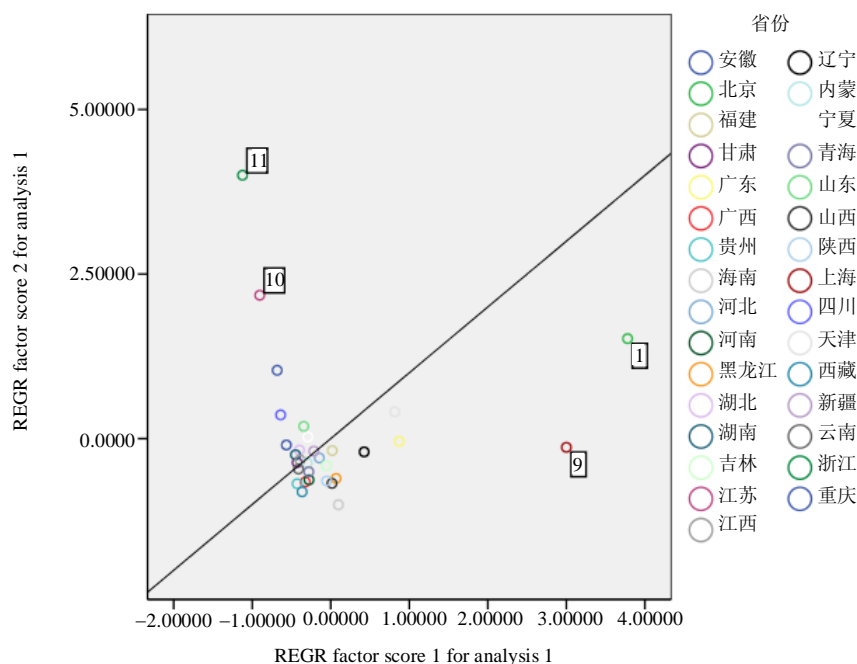


图 14-4 各省市两因子得分的散点图

执行以上操作后将各省市分为三类，结果如表 14-7 所示。

表 14-7 各省市因子得分及聚类情况

省 市	第一因子得分	第二因子得分	分类结果	省 市	第一因子得分	第二因子得分	分类结果
北京	3.77983	1.51859	1	湖北	-0.39973	-0.17441	3
天津	0.81383	0.40780	3	湖南	-0.44952	-0.24546	3
河北	-0.14581	-0.29445	3	广东	0.87085	-0.04330	3
山西	0.01489	-0.68057	3	广西	-0.32576	-0.65851	3
内蒙	-0.29983	-0.37042	3	海南	0.09679	-1.00483	3
辽宁	0.42466	-0.20229	3	重庆	-0.68314	1.03717	3
吉林	-0.05479	-0.41085	3	四川	-0.63881	0.35990	3
黑龙江	0.06702	-0.60460	3	贵州	-0.42883	-0.68617	3
上海	3.00024	-0.13269	1	云南	-0.41133	-0.46291	3
江苏	0.90495	2.17826	2	西藏	-0.36517	-0.80843	3
浙江	-1.12429	4.00113	2	陕西	-0.05308	-0.64398	3
安徽	0.56645	0.09701	3	甘肃	-0.42876	-0.36373	3
福建	0.01809	-0.18283	3	青海	-0.28111	-0.50374	3
江西	0.39323	-0.32454	3	宁夏	-0.29241	0.02442	3
山东	0.34529	0.18625	3	新疆	-0.21637	-0.19048	3
河南	0.27755	0.62732	3				

表 14-7 中的第一因子得分可以解释为各省市对房地产的需求水平。其取值越大,则说明房地产市场的需求越旺盛。从表中数据可知,排名最为靠前的是北京、上海、浙江三地,其中北京、上海的需求尤为旺盛。第二因子得分可以解释为各省市房地产市场的供给能力。其值越大,说明供给能力越强。从表中数据可知,浙江、江苏、北京三地的供给能力最强。

但是更多的时候需要了解的是各地房地产市场供需是否平衡的问题。根据聚类分析的结果,将所有省市大致分为三类:第一类{北京、上海},第二类{江苏、浙江},第三类{其余省市}。根据分类结果结合如图 14-4 所示的散点图,可以得出以下结论:目前北京、上海两地房地产市场存在严重的供不应求情况;江苏、浙江两地房地产市场供应已大于需求;而其余地区可以基本认为供需平衡。因此,国家可以根据各地房地产市场不同的特征制定相应的宏观调控政策。

14.1.3 模型的验证

为了说明模型分析的科学性,现在从《中国 2003 房地产发展报告——房地产蓝皮书》中选取了北京、广州、杭州三个城市的报告,来验证上节结论的合理性。

北京“房地产业目前处于高速发展期,运行态势总体良好,全市呈现供求两盛的局面。受北京 2008 年举办奥运会和首都率先实现现代化要求的影响,北京市房地产业依然存在较大上升空间。”“尽管 2003 年商品住宅投资额增幅减小,但是从 2001 年至 2003 年形成的 1700 多亿元的投资规模,将在几年内转变为巨大的建设规模和销售规模。”这说明北京市场供需两旺,供不应求,与上节的分析结论相吻合。

“2003 年 1~10 月,广州市生产总值达到 27413.32 亿元,同比增长 14.9%。”“据预测,未来几年,广州国民经济将继续保持 10% 以上的稳定增长速度,住宅建设和房地产业将继续保持稳定的增长势头。”“大量农村人口向城镇迁移,对城市基础设施建设提出了新的需求,也为城市住房建设提供了新的市场机制。”这也与上节的结论相一致。

“从 1999 年开始,杭州市商品房价持续上升且平均涨幅将近 20%,超过了国际惯例认可的 10% 的合理上限。由此也就形成了目前我国社会各界,特别是房地产行业强烈关注的杭州现象。”“2002 年房地产投资 196 亿元,同比增长 39.4%。2003 年 1~2 月投资额将近 14 亿元,同比增长 24%。”“从浙江省住宅与房地产业会上传出消息:今后 5 年,浙江省计划房地产投资将每年保持在 600 亿元以上,房地产投资要占全社会固定资产投资 20% 以上。”可见杭州属于投资过热城市,这说明模型中将浙江划为供大于求的结论是合理的。

从上面的 3 个个例来看,本节的模型是客观合理的。各地区的政府可以结合自身的具体情况制定相关政策,使我国房地产业和谐健康的发展。

14.2 SPSS在生物模型中的应用

从 20 世纪开始,随着数量遗传学和数量生态学等新兴学科的诞生。统计学在生物学中发挥了越来越大的作用。统计学在生物学中应用的一种主要形式就是生物数学模型。生物数学模型研究如何利用数学语言和数学工具,描述生物学的现象和规律。然后再将这些

结论用来解释、预测生物学现象和发现新规律。本节介绍一个利用 SPSS 软件估计生物数学模型中参数的应用实例。

14.2.1 问题描述

假设有一个生态系统，含有两种生物 A、B。其中 A 生物是捕食者，B 生物是被捕食者。假设 t 时刻捕食者 A 的数目为 $x(t)$ ，被捕食者 B 数目为 $y(t)$ ，它们之间满足以下变化规律：

$$\begin{cases} x'(t) = x(t)[\alpha_1 + \alpha_2 y(t)] \\ y'(t) = y(t)[\alpha_3 + \alpha_4 x(t)] \end{cases} \quad (14.1)$$

初始条件为：

$$\begin{cases} x(t_0) = \alpha_5 \\ y(t_0) = \alpha_6 \end{cases} \quad (14.2)$$

其中 $\alpha_k (1 \leq k \leq 6)$ 为模型的待定参数。通过对此生态系统的观测，可以得到相关的观测数据。观测数据的格式依次为：观测时刻 t_j 、A 生物数目 $x(t_j)$ 、B 生物数目 $y(t_j)$ 。

请利用有关数据，解决以下问题：

在观测资料有误差（时间变量不含有误差）的情况下，请分别利用观测数据 DATA2.TXT 和 DATA3.TXT，确定参数 $\alpha_k (1 \leq k \leq 6)$ 在某种意义下的最优解，并与仿真结果比较，进而改进数学模型。（2006 年全国研究生数学建模竞赛 B 题）

14.2.2 问题建模

14.2.1 节中的问题是 2006 年全国研究生数学建模竞赛 B 题中的一个最主要的问题。参赛选手在处理这个问题的时候提出了多种方法，也应用了多种工具来解决这个问题。下面介绍一下如何利用 SPSS 软件来分析这个问题。

1. 导入数据文件

由于题目中提供的数据文件是“*.txt”格式，因此首先要将其导入到 SPSS 中。执行以下操作：

执行【File】/【Read Text Data】命令，弹出【Open File】对话框

选择文件“DATA2.TXT”，单击【打开】按钮，选择打开数据文件 DATA2.TXT

利用数据文件打开向导，将 DATA2.TXT 导入到 SPSS 中

将变量依次命名为 t、X、Y

执行【File】/【Save】命令，弹出【Save】对话框

将文件保存为“data2.sav”，单击【保存】按钮，将原始数据保存为 SPSS 数据文件

执行以上操作后，题目中所提供的文本文件“DATA2.TXT”就成功导入到 SPSS 中，并保存为数据文件“data2.sav”。变量依次命名为“t”、“X”、“Y”。用同样的方法可以将文本文件“DATA3.TXT”中的数据保存到数据文件“data3.sav”中。

2. 建立初步模型

(14.1) 式是著名的生态数学模型——Lotka-Volterra 模型。对该微分方程组化简有：

$$y = a_3 \ln x + a_4 x - a_1 \ln y + c \quad (14.3)$$

其中 c 为任意常数， $a_3 = \frac{\alpha_3}{\alpha_2}$ ， $a_4 = \frac{\alpha_4}{\alpha_2}$ ， $a_1 = \frac{\alpha_1}{\alpha_2}$ 。由于这里主要讲解的是 SPSS 在这个问题中的应用。对于微分方程组具体的化简过程在这里省略不讲了。

以数据文件“data2.sav”为例，要估计 (14.1) 中的参数值可以转化为估计式 (14.3) 中的参数 a_1 、 a_3 、 a_4 。在 (14.3) 式中，将 $\ln x$ 、 $\ln y$ 也可以看作一个新的变量。因此现在就需要在数据文件中生成两列新变量，执行以下操作：

执行【Transform】/【Compute Variable】命令，弹出【Compute Variable】对话框
【Target Variable】框：LNX 定义生成新变量“LNX”
【Numeric Expression】框：LN(X) 定义新变量的计算公式
单击【OK】按钮 定义完成

执行以上操作之后，生成一列新变量“LNX”。用同样的方法再生成一列新变量“LNY”。现在的目标是，拟合 a_1 、 a_3 、 a_4 的值 \hat{a}_1 、 \hat{a}_3 、 \hat{a}_4 ，

$$\text{s.t} \quad \min \sum_{i=1}^n [(\hat{a}_4 x_i + \hat{a}_3 \ln x_i - \hat{a}_1 \ln y_i) - y_i]^2 \quad (14.4)$$

此时，题目中微分方程组参数的估计问题已经转换为在误差平方和最小的情况下估计多元线性函数 (14.3) 式中的参数的问题。其本质就是一个多元线性回归问题。执行以下操作：

执行【Analyze】/【Regression】/【Linear】命令，弹出【Linear】对话框
【Dependent】框：Y 定义 Y 为因变量
【Independent】框：X、LNX、LNY 定义自变量
单击【OK】按钮 定义完成

执行以上操作，可生成一系列的回归分析结果的相关表格。从表 14-8 可以看出，用多元线性回归模型来拟合 (14.3) 式中的 a_1 、 a_3 、 a_4 是非常恰当的。表格具体的判别方法请参阅第 10 章线性回归部分。

表 14-8 模型拟合度检验

Model Summary				
Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	0.999 ^a	0.997	0.997	1.29739

a. Predictors: (Constant), LNY, X, LNX

表 14-9 是回归分析的结果。最后的拟合结果为 $\hat{a}_1 = -19.887$ 、 $\hat{a}_3 = 99.834$ 、 $\hat{a}_4 = -9.991$ 。即 $\hat{y} = 99.834 \ln x - 9.991x - 19.887 \ln y - 139.472$ 。

表 14-9 回归分析结果

Coefficients ^a					
Model		Unstandardized Coefficients		Standardized Coefficients	t
		B	Std. Error	Beta	
1	(Constant)	-139.472	1.343		-103.823
	X	-9.991	0.109	-2.185	-91.417
	LN X	99.834	1.071	2.236	93.210
	LN Y	19.887	0.087	1.081	228.501

a. Dependent Variable: Y

为了定量比较各类模型下系数 a_i 的拟合精度，定义 ϕ_1 （平均均方误差）来描述估计的准确度：

$$\phi_1 = \sum_{i=1}^n [(\hat{a}_4 x_i + \hat{a}_3 \ln x_i - \hat{a}_1 \ln y_i) - y_i]^2 / n$$

在 SPSS 中， ϕ_1 有两种求法。第一种方法是通过执行【Transform】/【Compute】命令生成一系列新变量 $[(\hat{a}_4 x_i + \hat{a}_3 \ln x_i - \hat{a}_1 \ln y_i) - y_i]^2$ ，求其均值，即为 ϕ_1 的值。第二种方法则是通过读取表 14-10 中的残差 (Residual) 的平方和 (Sum of Squares) 项。此时该项取值为 247.433，即为 $[(\hat{a}_4 x_i + \hat{a}_3 \ln x_i - \hat{a}_1 \ln y_i) - y_i]^2$ 的值，再除以样本数即得 ϕ_1 的值。在数据文件“data2.sav”中，共有 151 个样本，所以求得 $\phi_1 = 1.6386$ 。

表 14-10 方差分析表

ANOVA ^b					
Model		Sum of Squares	df	Mean Square	F
1	Regression	87955.498	3	29318.499	17418.113
	Residual	247.433	147	1.683	
	Total	88202.931	150		

对数据文件“data3.sav”执行同样的操作，其拟合结果如表 14-11 所示。

表 14-11 模型 1 下参数的拟合结果

	α_1 / α_2	α_3 / α_2	α_4 / α_2	ϕ_1
data2	-19.8871	99.8339	-9.9910	1.6386
data3	-18.3749	89.1609	-8.9335	33.0114

data2 和 data3 在模型 1 下的拟合结果如图 14-5 和图 14-6 所示。其中星号代表模型 1 拟合出来的 y 的估计值，曲线代表 y 的真实值。在具体拟合的过程中，data3 中的原始数据 y 有 3 项为负值，对负数计算自然对数无意义。此时 SPSS 会自动剔除这 3 组观测值。由于它们所占的比重很小，且在生态系统中这些数据是没有实际意义的，所以剔除这 3 组观测值对最后的结果没有影响。

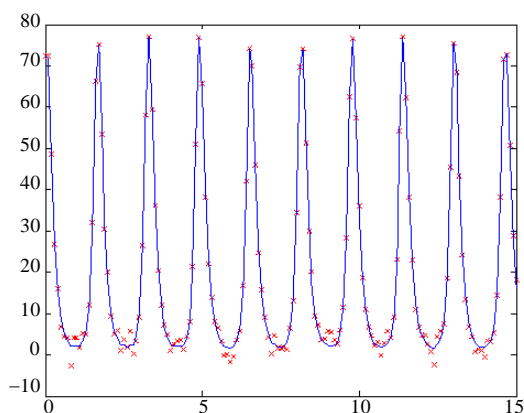


图 14-5 data2 在模型 1 下的拟合结果

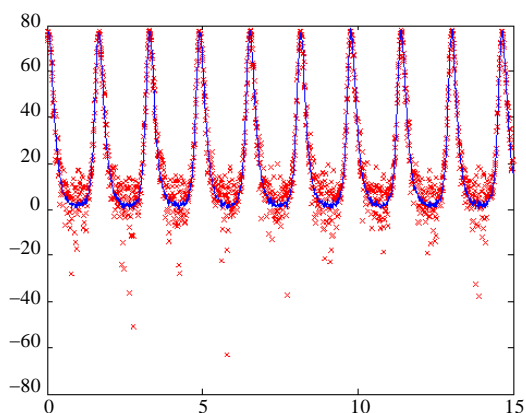


图 14-6 data3 在模型 1 下的拟合结果

3. 改进后的模型

通过分析表 14-11 中的数据和观察图 14-6 不难发现，对于 data3，即使在最小二乘的准则下，其拟合的结果也不是很好。这说明原始模型存在的问题，需要改进。

在模型 1 中，目标函数是极小化 y 的均方误差，由于在 (14.1) 式中， x 、 y 是一个平等的关系（纯粹只考虑数学意义下），所以现在一个自然的想法是尝试着极小化 x 的均方误差。

令 $b_2 = \frac{\alpha_2}{\alpha_4}$, $b_1 = \frac{\alpha_1}{\alpha_4}$, $b_3 = \frac{\alpha_3}{\alpha_4}$ ，则有

$$x = b_2 y + b_1 \ln y - b_3 \ln x + c$$

现在的目标是拟合 b_1 、 b_3 、 b_4 的估计值 \hat{b}_1 、 \hat{b}_2 、 \hat{b}_3

$$\text{s.t.} \quad \min \sum_{i=1}^n [(\hat{b}_2 y_i + \hat{b}_1 \ln y_i - \hat{b}_3 \ln x_i) - x_i]^2$$

此时仍是处理一个多元线性回归问题。对“data2.sav”执行以下操作：

执行【Analyze】/【Regression】/【Linear】命令，弹出【Linear】对话框

【Dependent】框：X

定义 X 为因变量

【Independent】框：Y、LNX、LNY

定义自变量

单击【OK】按钮

定义完成

执行以上操作之后模型的回归分析结果和方差分析表如表 14-12 和表 14-13 所示。此时的拟合结果为 $x = -0.098y + 1.961 \ln y + 9.986 \ln x - 13.917$ 。模型的残差平方和为 2.436。可见残差平方和大大降低。

表 14-12 回归分析结果

Coefficients^a

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	-13.917	0.055		-252.822	0.000
	Y	-.098	0.001	-.450	-91.417	0.000
	LNX	9.986	0.020	1.023	504.281	0.000
	LNY	1.961	0.020	0.487	98.653	0.000

表 14-13 方差分析表

ANOVA ^b						
Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	4215.526	3	1405.175	84797.084	0.000 ^a
	Residual	2.436	147	0.017		
	Total	4217.962	150			

令 $\phi_2 = \sum_{i=1}^n [(\hat{b}_2 y_i + \hat{b}_1 \ln y_i - \hat{b}_3 \ln x_i) - x_i]^2 / n$ 。在数据文件“data2.sav”中 $n=151$ ，利用残差平方和除以样本个数可以得出 $\phi_2 = 0.01624$ 。对数据文件“data3.sav”执行同样的操作，拟合结果如表 14-14 所示。

表 14-14 模型 2 下参数的拟合结果

	α_1/α_4	α_2/α_4	α_3/α_4	ϕ_2
data2	1.9612	-0.0984	-9.9862	0.01624
data3	1.5465	-0.0796	-9.8722	0.29389

data2 和 data3 在模型 2 下的拟合结果如图 14-7 和图 14-8 所示。其中星号代表模型 1 拟合出来的 x 的估计值，曲线代表 x 的真实值。

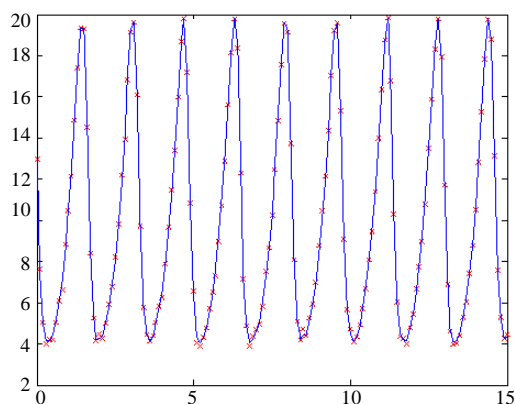


图 14-7 data2 在模型 2 下的拟合结果

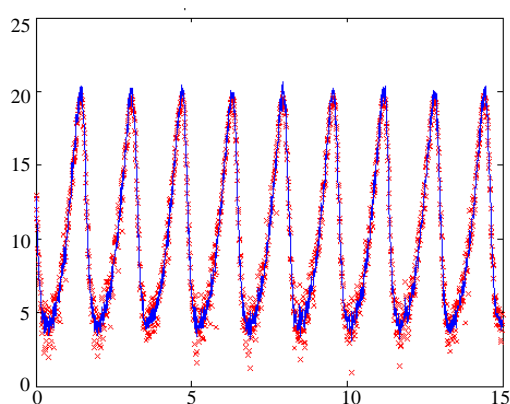


图 14-8 data3 在模型 2 下的拟合结果

比较表 14-14 中的 ϕ_2 和表 14-11 中的 ϕ_1 可以发现，对于相同的数据样本，拟合误差已经是数量级的下降了。同时比较图 14-7、图 14-8 与前面的图 14-5、图 14-6 也可以发现模型 2 的改进效果十分明显。

14.2.3 模型的讨论

从 (14.1) 式看， x 和 y 在数学意义上是完全平等的，但是为什么拟合 x 就比拟合 y 数量级的降低了误差，这就要回到模型的生态学背景了。

在式 (14.1) 中， x 表捕食者数目， y 表被捕食者数目，根据生活常识可知，捕食者数

目肯定小于被捕食者数目，且其扰动会极大地影响生态模型。举个简单的例子，一个森林中有 2 只老虎，100 只兔子，兔子数目不变，老虎从 2 只变为 1 只对生态系统的影响，肯定大于老虎数目不变，兔子数目从 100 只变成 99 只对生态系统的影响。所以，在有限的条件下，拟合模型捕食者的数目的效果比拟合被捕食者数目的效果要好。（参考文献：《统计方法在高精度参数估计问题中的应用》，罗应婷，吴荣军，杨钰娟）

14.3 SPSS在工程问题中的应用

在工程问题中，会涉及各种大批量数据的统计问题。比如在航空航天问题中，根据雷达的观测值，要估计飞行器运动状态的参数；在产品质量控制过程中，需要做产品质量可靠性分析，等等。若采用统计软件来处理这些问题，必将大大简化问题的复杂性。本节给出一个 SPSS 在具体工程问题中的应用实例。

14.3.1 问题描述

在泥石流的防治工程中，一次泥石流的总流量是十分重要的参数。然而，在一次具体的泥石流过程中，由于相当一部分泥石流汇入主河，所以泥石流总流量的计算十分困难。并且，一次泥石流之后，所能测得的数据也相当有限，仅能测得该次泥石流的最大流量和通过访问了解该次泥石流的总历时。因此，现在的目的就是寻找一种方法使之能够通过最大流量和总历时来估计一次泥石流的总流量。

根据泥石流暴涨暴落的特点，之前人们将泥石流的过程曲线概化为五边形（如图 14-9 所示），一次泥石流的总流量按下式计算：

$$W = \frac{19}{72} Q_c T_{\text{总}}$$

其中， Q_c 表示一次泥石流的最大流量， $T_{\text{总}}$ 表示一次泥石流的总历时， W 表示该次泥石流的总流量。

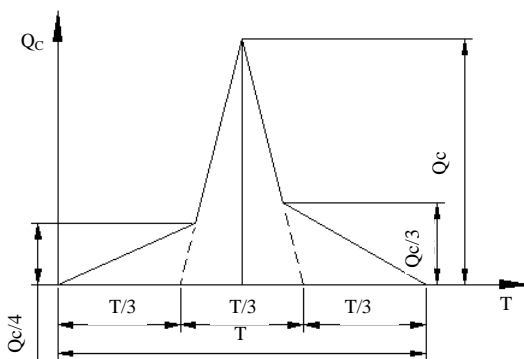


图 14-9 原始的五边形模型

这种方法由于过分简单，而显得粗糙。通过实际数据检验可知，这种方法的偏差过大。

本节将利用 SPSS 软件分析云南蒋家沟 7 年（1987~1993 年）共 57 次泥石流观测数据（数据来源：《云南蒋家沟泥石流运动观测资料集》，科学出版社），提出了一种由最大流量和总历时估计泥石流总流量的新方法，并且通过实际数据检验可知，其计算效果远远优于之前的五边形概化方法。

14.3.2 问题建模

下面将详细介绍如何通过 SPSS 软件来分析这个问题。

1. 数据文件的说明

在具体讲解本节问题之前，有必要对观测资料做简单介绍。观测资料共包括 57 次泥石流比较完整的观测数据。在实际过程中，泥石流都是阵发性的。因此，每次泥石流的观测资料中又包含了若干阵。如图 14-10 所示，即为一次泥石流的观测数据表。表中的各记录值为该次泥石流的各阵观测数据。表头编号“8901”代表该次泥石流是 1989 年发生的第一次泥石流。

泥石流运动要素数据表														
Date of debris flow kinematic observation										编号：(Code):8901				
序号	流态	龙头时间	龙尾时间	历时	泥面宽	泥深	测速距离	测速时间	流速	流量	容重	输沙率	径流量	备注
No	Type	T1 (hms)	T2 (hms)	T (s)	B (m)	H (m)	L (m)	t (s)	V (m/s)	Q (m ³ /s)	γ c (t/m ³)	Qc (t/s)	Wc (m ³)	Note
1	阵性流S	23:32:00	23:48:40	1000	20	0.8	50	6.3	7.94	127	2	2.095	63500	
2	阵性流S	1:00:52	1:01:22	30	25	1.9	72	6.1	11.8	560	2	900.11	8408	

图 14-10 泥石流运动要素数据表

因为共有 57 次泥石流，所以需要建立 57 个“*.sav”数据文件来保存各次泥石流的观测数据。比如 1989 年发生的第一次泥石流将保存在“8901.sav”中。

注意 对于问题的工程背景不了解的读者可能会混淆泥石流“次”与“阵”的概念。或许有的读者还会错误地认为本例就只有 57 个观测值。所以此处再次强调：每次泥石流是由若干阵组成的。比如每次泥石流平均有 100 阵，那么本例就一共有 5700 个数据样本。

由于泥石流按流态分为阵性流和连续流两种形态，而这里仅关心其在阵性流情况下的问题，所以首先要对数据文件作拆分。以数据文件“8905.sav”为例，执行以下操作：

执行【Data】/【Split File】命令

选择“Compare Groups”单选框

【Groups based on】：流态

单击【OK】按钮

弹出【Split File】对话框

定义按照变量拆分文件

定义按照流态拆分文件

定义完成

执行以上操作之后，数据文件“8905.sav”就按流态分为两组。虽然在表面上可能看不出区别，但是在后面的分析中，都是将阵性流和连续流分别处理了。

2. 由最大流量模拟泥石流的流量分布函数

首先模拟一次泥石流的流量分布函数。仍以 1989 年第 5 次泥石流数据为例，对数据文件“8905.sav”执行以下操作：

执行【Graphs】/【Legacy Dialogs】/【Histogram】命令，弹出【Histogram】对话框	
【variables】：流量	定义绘制流量的直方图
选中“Display normal curve”复选框	定义在图形上绘制相应的正态曲线

执行以上操作之后，对于阵性流，生成如图 14-11 所示的流量直方图。

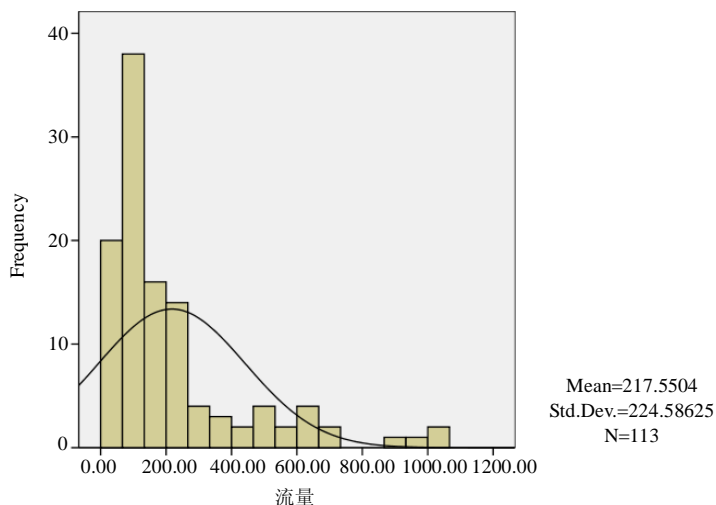


图 14-11 数据文件“8905.sav”的流量直方图

通过观察 1987~1993 年共 57 次泥石流的流量直方图可以发现，大部分泥石流的流量分布具有正偏的特点（如图 14-11 所示），但是也有部分泥石流由于阵次较少，呈现出负偏的特点。基于以上原因，选取 Weibull 分布来模拟一次泥石流的分布。Weibull 分布的概率密度函数如下：

$$f(x) = crx^{r-1}e^{-cx^r} \quad (c > 0, r > 0) \quad (14.5)$$

选取 Weibull 分布主要是因为当 r 取较小值时，分布的密度函数是为正偏的；而当 r 取较大值时，分布的密度函数为负偏。对于接近 3.5 的值，近似对称。这正好与用实际数据拟合出来的泥石流的分布特征相吻合。但是这只是一种主观地判断，可以进一步通过 Weibull 分布的 P-P 图来检验分布的拟合程度。仍以 1989 年第 5 次泥石流数据文件“8905.sav”为例，执行以下操作：

执行【Analyze】/【Descriptive Statistics】/【P-P Plots】命令，弹出【P-P】对话框	
【variables】：流量	定义绘制流量的 P-P 图
【Test Distribution】下拉列表：Weibull	定义检验的分布类型为 Weibull 分布
单击【OK】按钮	定义完成

执行以上操作后，做出流量关于 Weibull 分布的 P-P 图和 P-P 趋势图。分析图形可以

发现,其流量数据的确很好地服从于 Weibull 分布。数据的具体分析标准请参阅第 4 章 P-P 图的相应部分。

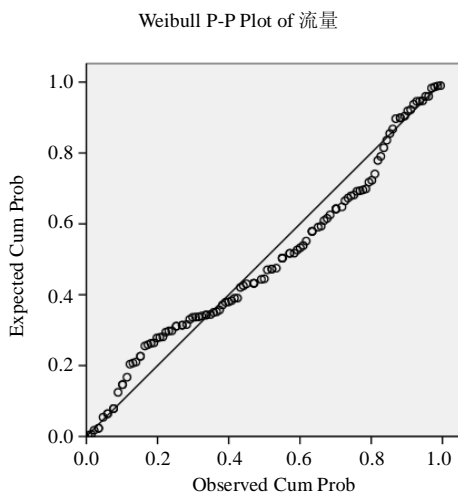


图 14-12 数据文件“8905.sav”的流量 P-P 图

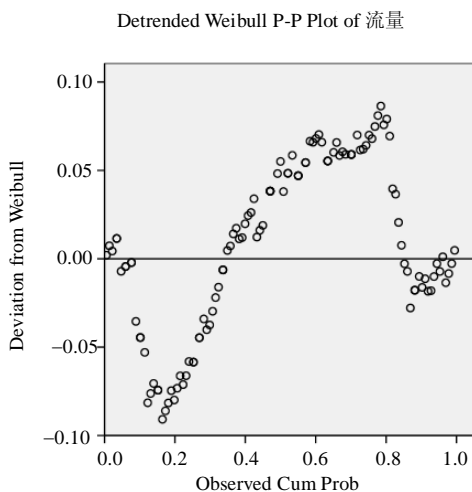


图 14-13 数据文件“8905.sav”的流量 P-P 趋势图

现在的任务即为估计 Weibull 分布中的参数。由 (14.5) 式计算可知, Weibull 分布的分布函数如下:

$$F(x) = 1 - e^{-cx^r} \quad (14.6)$$

由于 Weibull 分布是双参数模型,因此可以选择两个流量分布函数的分位点来估计其参数值。设 $Q_{1/2}$ 和 $Q_{3/4}$ 分别表示一次泥石流流量的 1/2 分位点和 3/4 分位点,代入 (14.6) 式有:

$$1 - e^{-cQ_{1/2}^r} = 1/2 \quad (14.7)$$

$$1 - e^{-cQ_{3/4}^r} = 3/4 \quad (14.8)$$

由 (14.7)、(14.8) 式可以推断出:

$$r = \frac{\ln \frac{1}{2}}{\ln(\frac{Q_{1/2}}{Q_{3/4}})}, \quad c = \frac{-\ln(\frac{1}{2})}{Q_{1/2}^r}$$

这样,就通过流量的 1/2 和 3/4 分位点确定了 Weibull 分布的参数值。此处之所以选取这两个分位点而非其他,主要是因为位于中部的分位点比位于边缘位置的分位点具有更好的稳定性。然而,在实际问题中,仅能测到最大流量。因此,还需要寻找最大流量与流量分位点之间的关系。这种关系可能是线性的,也可能是非线性的。需要根据已有的数据选择适当的模型并估计模型参数。

已知有 57 次泥石流的观测资料,仍以数据文件“8905.sav”为例,执行以下操作:

执行【Analyze】/【Descriptive Statistics】/【Frequencies】命令,弹出【Frequencies】对话框
【variables】: 流量 选择分析变量“流量”

单击【Statistics】按钮	弹出【Statistics】对话框
【Statistics】对话框：	
选中“Quartiles”复选框	定义计算流量的四分位点
选中“Maximum”复选框	定义计算流量的最大值
单击【Continue】按钮	【Statistics】对话框定义完成
单击【OK】按钮	定义完成

执行以上操作之后，生成表 14-15。由表 14-15 可知，对于数据文件“8905.sav”，其流量的最大值为 1050，1/2 分位点为 131，3/4 分位点为 249.85。

表 14-15

Statistics			
流量			
连续流 C	N	Valid	6
		Missing	0
	Maximum		10.00
	Percentiles	25	0.8750
		50	3.5000
		75	5.5000
阵性流 S	N	Valid	113
		Missing	1
	Maximum		1050.00
	Percentiles	25	77.8500
		50	131.0000
		75	249.8500

对 57 次泥石流的数据文件分别执行以上操作，则可求出 1987~1993 年各次泥石流流量的 1/2 和 3/4 分位点及流量最大值，将其保存在一个新的数据文件“87-93.sav”中。变量名分别为“二一”、“四三”和“极大”。下面要寻找流量分位点与最大值之间的关系。对数据文件“87-93.sav”执行以下操作：

执行【Graph】/【Legacy Dialogs】/【Scatter】命令，选择【Simple Scatter】		定义 绘制简单散点图
【Simple Scatter】对话框：		
【Y Axis】框：二一		定义简单散点图 Y 轴
【X Axis】框：极大		定义简单散点图 X 轴
单击【OK】按钮		定义完成

执行以上操作之后，生成如图 14-14 所示的流量最大值与流量 1/2 分位点之间对应关系的散点图。同样的方法可以得到如图 14-15 所示的流量最大值与流量 3/4 分位点之间对应关系的散点图。

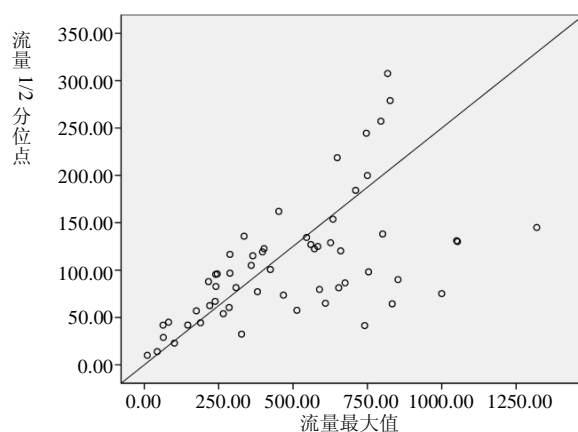


图 14-14 流量最大值与 1/2 分位点关系散点图

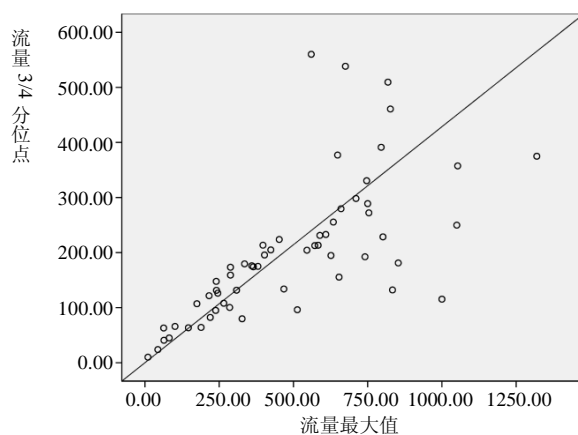


图 14-15 流量最大值与 3/4 分位点关系散点图

观察散点图，发现可以用线性回归的方法来拟合流量分位点和最大值之间的关系。为了便于给出函数的物理解释，下面用不带常数项的线性回归模型拟合二者的关系。执行以下操作：

执行【Analyze】/【Regression】/【Linear】命令，弹出【Linear】对话框	
【Dependent】框：二一	定义流量 1/2 分位点为因变量
【Independent】框：极大	定义极大流量
单击【Options】按钮	弹出【Options】对话框
【Options】对话框：	
去掉“Include constant in equation”复选框的默认勾选	定义拟合模型中不包括常数项
单击【Continue】按钮	【Options】对话框定义完成
单击【OK】按钮	定义完成

执行以上操作后，生成表 14-16。

表 14-16 回归分析结果

Coefficients ^{a,b}					
Model	Unstandardized Coefficients		Standardized Coefficients	t	Sig.
	B	Std. Error	Beta		
1 极大	0.192	0.013	0.889	14.653	0.000

从表 14-16 可知：

$$Q_{1/2} = 0.192Q_c$$

$$Q_{3/4} = 0.376Q_c$$

这样，就给出了一种通过一场泥石流的最大流量来模拟其分布的方法。

3. 计算一次泥石流的总流量

下面将要讨论一次泥石流总量计算的问题。在一次泥石流过程中，对于第 i 阵阵性流，有

$$W_i = \frac{1}{2}Q_iT_i$$

其中， W_i 为第 i 阵的总流量， Q_i 为第 i 阵的流量， T_i 表第 i 阵的总历时。则对于一场 N 阵的泥石流，其总流量 W 有

$$W = \sum_{i=1}^N W_i = \sum_{i=1}^N \frac{1}{2}Q_iT_i \quad (14.9)$$

此时，一个自然的想法就是用流量的平均值来代替各阵流量，用实际流动的总历时来代替各阵的历时之和，则（14.9）式可以近似地表为

$$W \approx \frac{1}{2}\bar{Q}T \quad (14.10)$$

其中 \bar{Q} 表示流量的平均值， T 表示实际流动的总时间。

由于是用 Weibull 分布来模拟流量 Q 的分布，所以流量 Q 的平均值为：

$$\bar{Q} = \int_0^{\infty} xf(x)dx = \frac{\Gamma(1+1/r)}{c^{1/r}} \quad (14.11)$$

其中 $\Gamma(s)$ 表示 gamma 函数。 c 、 r 从前面的分析可知，可以由最大流量计算出来。

对于泥石流，由于阵与阵之间的断流现象十分普遍。因此还需要讨论实际流动的总时间与总历时（含间断时间）之间的关系。由于阵与阵之间的间断时间是随机的，为了简化问题，此处用阵与阵之间的期望均值来描述。根据蒋家沟的历史数据，用 SPSS 执行【Analyze】/【Regression】/【Linear】命令，拟合不带常数项的线性回归模型，有

$$T = 0.2656T_{\text{总}} \quad (14.12)$$

将（14.11）、（14.12）式代入（14.10）式，给出一个通过一次泥石流的最大流量和总历时计算其总流量的方法。

14.3.3 模型的检验

下面分别用以前概化的方法及 14.1.2 节所提出的方法来计算云南蒋家沟 1994 年各场

泥石流的总流量，计算结果如表 14-17 所示。

表 14-17 两种方法计算结果的比较

序 号	极大流量 (m^3/s)	总 历 时 (s)	实际流量 (m^3)	以前方法计算的总流量 (m^3)	新方法计算的总流量 (m^3)
9401	186.7	14820	47873	730152	99327
9402	1382.5	31500	1310950	11528513	1563300
9403	2027.8	17880	608583	9567836	1301500
9404	929.2	13920	338307	3413261	464310
9405	585.0	11820	192885	1824712	248220
9406	288.4	9960	228891	758011	103120

通过实际计算可知，14.1.2 节中方法的效果优于之前概化的方法。

但是，分析结果同实际数据仍然有一定偏差，主要是由于以下几点原因：

首先，仅用已知的两个量来计算复杂的泥石流过程的总流量，这本身就是一个很困难的问题。

其次，在实际问题中，常常出现这样的情况，即一场泥石流的最大流量和总历时均大于另一场，但是其径流量小于另一场，如表 14-1 中的 9405 和 9406。因此，泥石流过程本身的不稳定性造成了误差的不可避免。

第三，在具体的拟合过程中，基本都是采用线性的模型来拟合问题。如果采用更复杂的模型来拟合的话势必进一步提高模型精度。

从本节的例子可以看出，在做实际问题的时候，贯穿在其中的其实更多的是实际工作者对问题的掌控与理解。可能不同的人对同一个问题有不同的解决方法。但是在具体实现方法的时候，SPSS 就是一个必不可少的工具了。本例中分布模型的确定与检验，模型具体参数的拟合等都是通过 SPSS 计算出来的。（参考文献：《从突发性流体最大流量估计总流量的统计方法》，罗应婷，陈宁生，朱允民）

14.4 SPSS在证券分析中的应用

在国外的证券行业中，统计分析是一个成熟和必需的工具。近年来伴随着中国证券市场的发展，越来越多的学者将统计方法引入到证券业之中。而 SPSS 也是这些工作者经常使用的工具之一。本节将通过中国股票市场的日历效应分析，介绍 SPSS 在证券行业中的应用。

14.4.1 问题描述

在股票市场上，收益、风险等指标一般都有随着日历变化的特征，称之为日历效应。日历效应可以分为周效应——指标在一周内各日表现出不同的特征；月效应——指标在不同月份具有不同的特征。国内外许多学者对价格变动的日历效应进行了大量研究，发现收益率和交易量都存在显著的日历效应。Rozeff 和 Kinney（1976 年）发现股票市场指数收益

率在 1 月份更高一些；Keim（1980 年）对这一现象用不同规模的股票组合进行研究，发现 1 月份的效应和绝大多数小规模股票有着密切的联系。Frech（1980 年）注意到股票收益率的周内效应——周日的收益率更低一些；Frech 和 Roll 发现方差也具有日历效应等。

下面以我国沪深股市为代表，考察是否也存在周效应，分析其具体特征及沪深股市的异同。本节中采用的数据样本为沪深股市 2005 年 9 月 5 日~2006 年 12 月 5 日的交易数据。（数据来源：Wind 资讯）

14.4.2 问题建模

下面详细介绍如何通过 SPSS 软件分析我国沪深股市的日历效应。

1. 创建 SPSS 数据文件并整理数据

首先将 2005 年 9 月 5 日~2006 年 12 月 5 日沪深股市共 606 条交易数据保存在数据文件“zhengquan.sav”中。该数据文件的变量名及标签如图 14-16 所示。

Name	Label
JYRQ	交易日期
JYS	交易所
KP	开盘指数
ZG	日最高指数
ZD	日最低指数
SP	收盘指数
ZDS	涨跌数
ZDFD	涨跌幅度（%）
CJL	成交量
CJE	成交额

图 14-16 数据文件“zhengquan.sav”的变量名及其标签

由于 14.4.1 节中要考察的问题是沪深股市是否具有周效应。而数据文件“zhengquan.sav”中的交易日期是 yy/mm/dd 格式的日期型变量，因此首先要推算出各日期所对应的星期数。

对于数据文件“zhengquan.sav”执行以下操作：

执行【Transform】/【Compute Variable】命令，弹出【Compute Variable】对话框	
【Target Variable】框：XQ	定义生成新变量“XQ”
【Numeric Expression】框：XDATE.WKDAY(JYRQ)-1	定义新变量的计算公式
单击【Type&Label】按钮	
【Type&Label】对话框：	
【Label】框：星期	定义新变量的变量标签
单击【Continue】按钮	变量标签定义完成
单击【OK】按钮	定义完成

执行以上操作后，生成新变量“XQ”，它代表了各交易日所对应的星期数。需要注意的是，函数 XDATE.WKDAY() 可以将日期转化为其对应的星期数。但是在国外由于通常把星期日作为一周的第一天。因此，SPSS 的转换准则是将星期日赋值为 1，星期一赋值为

2, ……，依次类推。所以，为了和国内通常的标准一致，这里在作转换的时候，将函数 XDATA.WKDAY() 所得的值全部减 1。对于星期日，SPSS 下取值为 1，减 1 之后就变成 0 了。但是因为我国证券市场只在星期一到星期五进行交易，所以这种特殊的情况可以不予考虑。

2. 计算描述性统计量

选取沪深股市的指数每日涨跌幅度和成交量两项指标来研究其周效应。对于数据文件“zhengquan.sav”执行以下操作：

执行【Data】/【Split File】命令，弹出【Split File】对话框	
选择“Compare Groups”单选框	定义按照变量拆分文件
【Groups based on】：JYS、XQ	定义按照交易所和星期拆分文件
单击【OK】按钮	文件拆分定义完成
执行【Analyze】/【Descriptive Statistics】/【Frequencies】命令	
弹出【Frequencies】对话框	
【Variables】：ZDFD、CJL	选择指数涨跌幅度和成交量两项指标
单击【Statistics】按钮	弹出【Statistics】对话框
【Statistics】对话框：	
选中“Mean”复选框	定义计算变量的均值
选中“Std.deviation”复选框	定义计算变量的标准差
单击【Continue】按钮	【Statistics】对话框定义完成
单击【OK】按钮	描述性统计量计算定义完成

执行以上操作之后，计算出沪深股市按星期划分的描述性统计量如表 14-18 所示。

表 14-18 沪深股市周效应分析表

	沪 市				深 市			
	日涨跌幅 均值 (%)	日涨跌幅 度标准差	日成交量均 值 (万股)	日成交量 标准差	日涨跌幅 度均值 (%)	日涨跌幅 度标准差	日成交量均 值 (万股)	日成交量 标准差
星期一	0.4213	1.31157	337240	183844	0.4705	1.60792	42190.27	24887.43
星期二	0.1089	1.18174	343049.5	177644.5	0.1426	1.30311	43540.99	27784.64
星期三	0.1910	1.19435	332493.1	151707	0.1793	1.4593	42425.9	24264.91
星期四	0.0262	1.24009	339727.7	165527.4	0.0788	1.35941	42780.48	23397.27
星期五	0.2805	1.17372	342171.5	177798.7	0.3002	1.17326	41515.59	21897.31

对数据文件“zhengquan.sav”执行以下操作，将表格中的数据用图形直观表示出来。

执行【Data】/【Split File】命令，弹出【Split File】对话框	
选择“Analyze all cases”单选框	选择取消开始定义的文件拆分
单击【OK】按钮	文件拆分取消完成完成
执行【Graphs】/【Legacy Dialogs】/【Bars】命令，弹出【Bar Charts】定义对话框	
选择“Clustered”、“Summaries for groups of cases”	弹出【Clustered Bar】对话框
【Bars Represent】组：	

选择“Other statistic”单选框	选择自定义条图长条的统计意义
【Variables】框：ZDFD	定义长条代表涨跌幅度
单击【Change statistic】按钮	弹出【statistic】对话框
【statistic】对话框：选中“Mean of Values”	定义长条代表的是涨跌幅度的均值
单击【Continue】按钮	【Statistics】对话框定义完成
【Category Axis】：XQ	选择“星期”作为分类变量
【Define Clustered by】：JYS	选择“交易所”作为分组变量
单击【OK】按钮	涨跌幅度的均值条图定义完成

执行以上操作后，可以绘制出如图 14-17 所示的沪深股市涨跌幅度的均值周效应比较图。用类似的方法绘制出沪深股市涨跌幅度的方差周效应比较图、成交量均值周效应比较图、成交量方差周效应比较图，如图 14-18～图 14-20 所示。

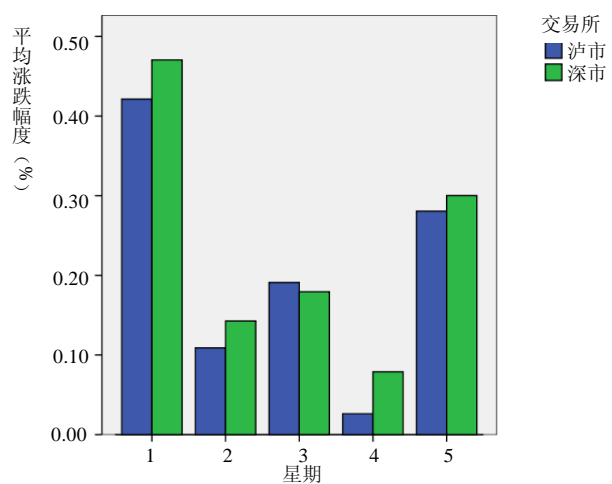


图 14-17 沪深股市涨跌幅度的均值周效应比较图

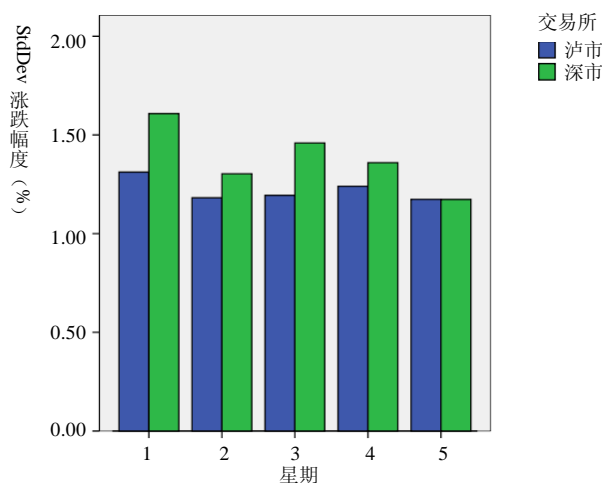


图 14-18 沪深股市涨跌幅度的方差周效应比较图

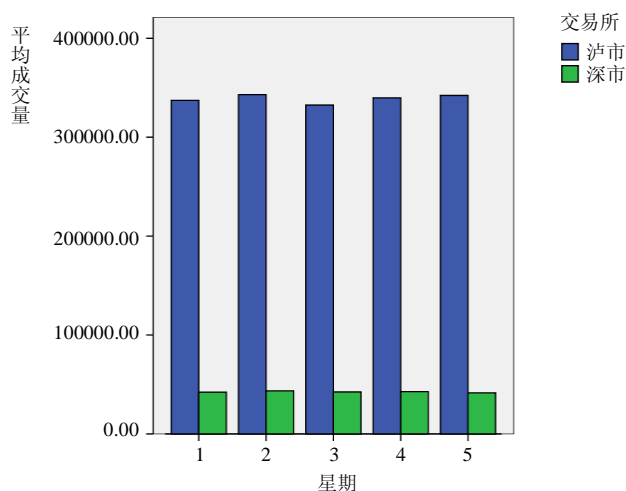


图 14-19 沪深股市成交量的均值周效应比较图

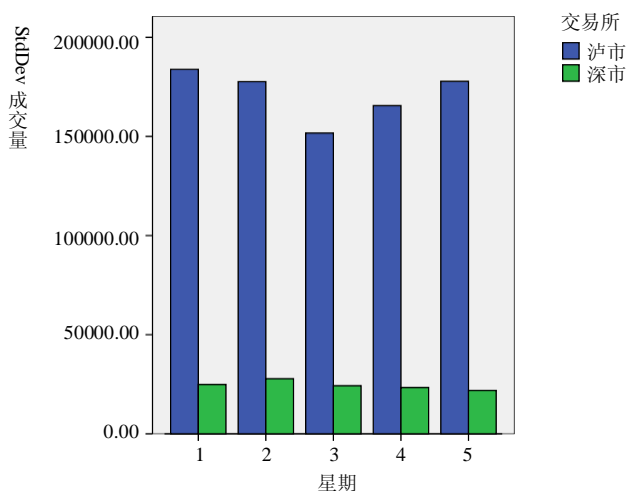


图 14-20 沪深股市成交量的方差周效应比较图

从表 14-18 的数据和图 14-17~图 14-20 可以大致得出以下结论:

- 沪深股市在星期一时涨跌幅度均值和方差都大于一星期内其他几日。
- 沪深股市的成交量在一星期各日内没有显著的差异。
- 沪深股市的涨跌幅度之间没有明显差异。
- 沪深股市的成交量表现出明显的“沪强深弱”趋势。

3. 对模型的进一步检验

前面根据 SPSS 所计算的描述性统计量及绘制的图形得出了 4 条结论。但是这只是一种主观的判断,有必要对这些结论进行进一步的检验。

(1) 首先分别对沪深股市的涨跌幅度做单因素方差分析,考察我国股票市场是否真的存在周效应现象,执行以下操作:

执行【Data】/【Split File】命令，弹出【Split File】对话框	
选择“Compare Groups”单选框	定义按照变量拆分文件
【Groups based on】: JYS	定义按照交易所拆分文件
单击【OK】按钮	文件拆分定义完成
执行【Analyze】/【Compare Means】/【One-Way ANOVA】命令，弹出【One-Way ANOVA】对话框	
【Dependent List】: ZDFD	定义涨跌幅度为指标变量
【Factor】: XQ	定义星期为因素变量
单击【Post Hoc】按钮	弹出【Post Hoc】对话框
【Post Hoc】对话框：	
选中“LSD”复选框	定义用LSD法进行多重比较检验
单击【Continue】按钮	【Post Hoc】对话框定义完成
单击【Options】按钮	弹出【Options】对话框
【Options】对话框：	
选中“Homogeneity of variance test”复选框	定义检验各组的方差齐次性
选中“Means plot”复选框	定义绘制均值的图形
单击【Continue】按钮	【Options】对话框定义完成
单击【OK】按钮	定义完成

执行以上操作之后，得出表 14-19，从表中可以看出沪深股市各星期期间涨跌幅度的方差是齐次性的。具体的判别标准请参阅第 8 章方差分析部分。

表 14-19 方差齐次性检验结果

Test of Homogeneity of Variances

涨跌幅度 (%)

交易所	Levene Statistic	df1	df2	Sig.
沪市	0.337	4	297	0.853
深市	1.209	4	297	0.307

表 14-20 是方差分析表。从表格中可以看出，虽然前面的描述性统计量分析认为星期一涨跌幅度的均值大于其他各日，但是这种差异是不显著的。

表 14-20 方差分析表

ANOVA

涨跌幅度 (%)

交易所		Sum of Squares	df	Mean Square	F	Sig.
沪市	Between Groups	5.645	4	1.411	0.946	0.437
	Within Groups	442.883	297	1.491		
	Total	448.528	301			
深市	Between Groups	5.755	4	1.439	0.746	0.561
	Within Groups	572.446	297	1.927		
	Total	578.200	301			

图 14-21 和图 14-22 是沪深两市股票涨跌幅度均值的线图。结合图形和表 14-20 可以判断, 虽然股票市场在一星期内的各日涨跌幅度有所差异, 但是这种差异是不显著的。不能因此就片面地判断我国的证券市场的指数涨跌幅度存在周效应。

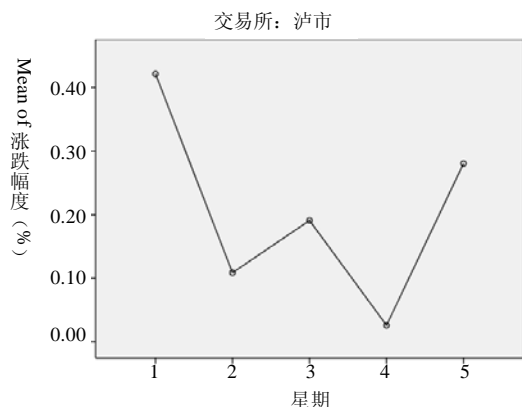


图 14-21 沪市涨跌幅度均值的线图

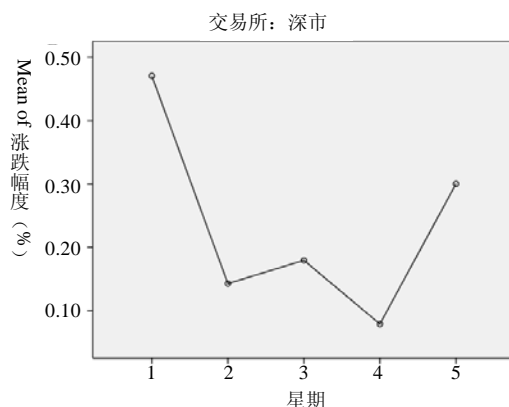


图 14-22 深市涨跌幅度均值的线图

(2) 对沪深股市的日成交量做同样的分析, 分析结果如下。

分析表 14-21 和图 14-23、图 14-24 可以得出以下结论: 沪深两市在一周内的日成交量虽然有一定差异, 但是这种差异是不显著的。不能认为我国的证券市场的成交量存在周效应。

表 14-21 方差分析表
ANOVA

成交量		Sum of Squares	df	Mean Square	F	Sig.
沪市	Between Groups	4.4E+009	4	1101325022	0.037	0.997
	Within Groups	8.8E+012	298	2.948E+010		
	Total	8.8E+012	302			
深市	Between Groups	1.4E+008	4	33784778.65	0.056	0.994
	Within Groups	1.8E+011	298	602006455.6		
	Total	1.8E+011	302			

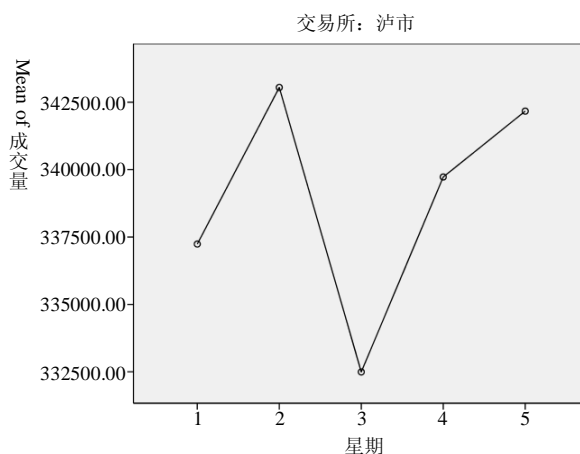


图 14-23 沪市日成交量均值的线图

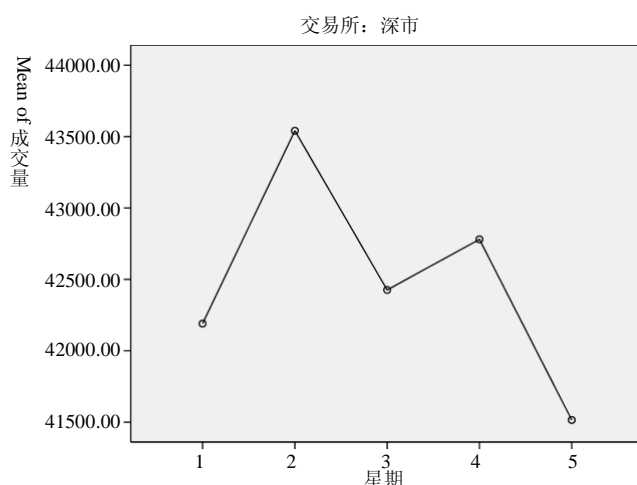


图 14-24 深市日成交量的线图

(3) 股值涨跌幅度和成交量的相关分析。

从图 14-21、图 14-22 可以看出，沪深两市的涨跌幅度的一周内的均值曲线以星期三为中心呈现出 W 型，而日成交量的均值曲线正好近似以星期三为中心呈现出 M 型。一个自然的想法就是寻找二者之间的相关性。执行以下操作：

执行【Analyze】/【Correlate】/【Bivariate】命令，弹出【Bivariate】对话框	
【Variables】：CJL、ZDFD	选择分析成交量与涨跌幅度间相
关系	
选中“Pearson”和“Spearman”复选框	计算两类相关系数
选中“Two-tailed”单选框	对相关系数进行双侧检验
选中“Flag significant correlations”选项	标识有统计意义的相关系数
单击【Options】按钮	弹出【Options】对话框
【Options】对话框：	
选中“Means and standard deviations”复选框，输出变量的均值和标准差	
单击【Continue】按钮	【Options】对话框定义完成
单击【OK】按钮	定义完成

执行以上操作之后，生成表 14-22 是沪深股市成交量和涨跌幅度的描述性统计量。

表 14-22 描述性统计量

Descriptive Statistics				
交易所		Mean	Std.Deviation	N
沪市	成交量	338923.2	170588.87054	303
	涨跌幅度 (%)	0.2052	1.22071	302
深市	成交量	42492.91	24381.96753	303
	涨跌幅度 (%)	0.2338	1.38598	302

表 14-23 是 Pearson 相关系数的结果，从表格中可以看出对于沪市，成交量和涨跌幅度的相关系数为 0.111，其 Sig.取值近似为 0.05。对于深市，相关系数为 0.142，其 Sig.取值为 0.013。说明成交量与涨跌幅度之间不存在相关关系的可能性很小，但是还是不能确定两者具体的相关程度。

表 14-23 Pearson 相关系数

Correlations				
交 易 所			成 交 量	涨跌幅度 (%)
沪市	成交量	Pearson Correlation	1	0.111
		Sig.(2-tailed)		0.055
		N	303	302
	涨跌幅度 (%)	Pearson Correlation	.111	1
		Sig.(2-tailed)	.055	
		N	302	302
深市	成交量	Pearson Correlation	1	0.142*
		Sig.(2-tailed)		0.013
		N	303	302
	涨跌幅度 (%)	Pearson Correlation	0.142*	1
		Sig.(2-tailed)	0.013	
		N	302	302

*. Correlation is significant at the 0.05 leve (2-tailed).

表 14-24 是 Spearman 相关系数的结果，从表 14-24 可以得出和上表结论类似的结果。

表 14-24 Spearman 相关系数

Correlations					
交 易 所				成 交 量	涨跌幅度（%）
沪市	Spearman's rho	成交量	Correlation Coefficient	1.000	0.128*
			Sig.(2-tailed)		0.027
			N	303	302
		涨跌幅度（%）	Correlation Coefficient	0.128*	1.000
			Sig.(2-tailed)	0.027	
			N	302	302
深市	Spearman's rho	成交量	Correlation Coefficient	1.000	0.214**
			Sig.(2-tailed)		0.000
			N	303	302
		涨跌幅度（%）	Correlation Coefficient	0.214**	1.000
			Sig.(2-tailed)	0.000	
			N	302	302

*. Correlation is significant at the 0.05 level (2-tailed).

**. Correlation is significant at the 0.01 level (2-tailed).

14.4.3 模型的讨论

从 14.4.2 节的模型可以看出，沪深两市的涨跌幅度和，成交量之间有应该一定的相关关系，但还有待进一步的分析。并且涨跌幅度大致呈 **W** 型；成交量大致呈 **M** 型。沪深两市在一周各天虽然涨跌幅度和成交量都各有差异，但是这种差异都不显著。因此，不能简单地判定我国的股票市场内存在着明显的周效应。还有待进一步的分析。

对于证券市场中的数据，统计方法还大有用武之地。在本节用 **SPSS** 软件对证券市场进行了简单地分析，希望能够起到抛砖引玉的作用。（参考文献：《中国股票市场的日历效应分析》，薛继锐，顾岚）



《SPSS 统计分析从基础到实践(第2版)》读者交流区

尊敬的读者:

感谢您选择我们出版的图书,您的支持与信任是我们持续上升的动力。为了使您能通过本书更透彻地了解相关领域,更深入的学习相关技术,我们将特别为您提供一系列后续的服务,包括:

1. 提供本书的修订和升级内容、相关配套资料;
2. 本书作者的见面会信息或网络视频的沟通活动;
3. 相关领域的培训优惠等。

请您抽出宝贵的时间将您的个人信息和需求反馈给我们,以便我们及时与您取得联系。

您可以任意选择以下三种方式与我们联系,我们都将记录和保存您的信息,并给您提供不定期的信息反馈。

1. 短信

您只需编写如下短信: B10010+您的需求+您的建议

发送到1066 6666 789 (本服务免费,短信资费按照相应电信运营商正常标准收取,无其他信息收费)

为保证我们对您的服务质量,如果您在发送短信24小时后,尚未收到我们的回复信息,请直接拨打电话 (010) 88254369。

2. 电子邮件

您可以发邮件至jsj@phei.com.cn**或**editor@broadview.com.cn**。**

3. 信件

您可以写信至如下地址: 北京万寿路173信箱博文视点, 邮编: 100036。

如果您选择第2种或第3种方式,您还可以告诉我们更多有关您个人的情况,及您对本书的意见、评论等,内容可以包括:

- (1) 您的姓名、职业、您关注的领域、您的电话、E-mail地址或通信地址;
- (2) 您了解新书信息的途径、影响您购买图书的因素;





电子工业出版社

PUBLISHING HOUSE OF ELECTRONICS INDUSTRY

Broadview[®]
www.broadview.com.cn

(3) 您对本书的意见、您读过的同领域的图书、您还希望增加的图书、您希望参加的培训等。

如果您在后期想退出读者俱乐部，停止接收后续资讯，只需发送“B10010+退订”至10666666789即可，

或者编写邮件“B10010+退订+手机号码+需退订的邮箱地址”发送至邮箱：market@broadview.com.cn 亦可

取消该项服务。

同时，我们非常欢迎您为本书撰写书评，将您的切身感受变成文字与广大书友共享。我们将挑选特别优秀的作品转载在我们的网站(www.broadview.com.cn)上，或推荐至CSDN.NET等专业网站上发表，被发表的书评的作者将获得价值50元的博文视点图书奖励。

我们期待您的消息！

博文视点愿与所有爱书的人一起，共同学习，共同进步！

通信地址：北京万寿路 173 信箱 博文视点 (100036) 电话：010-51260888

E-mail：jsj@phei.com.cn，editor@broadview.com.cn

反侵权盗版声明

电子工业出版社依法对本作品享有专有出版权。任何未经权利人书面许可，复制、销售或通过信息网络传播本作品的行为；歪曲、篡改、剽窃本作品的行为，均违反《中华人民共和国著作权法》，其行为人应承担相应的民事责任和行政责任，构成犯罪的，将被依法追究刑事责任。

为了维护市场秩序，保护权利人的合法权益，我社将依法查处和打击侵权盗版的单位和个人。欢迎社会各界人士积极举报侵权盗版行为，本社将奖励举报有功人员，并保证举报人的信息不被泄露。

举报电话：(010) 88254396; (010) 88258888

传 真：(010) 88254397

E-mail: dbqq@phei.com.cn

通信地址：北京市万寿路 173 信箱

电子工业出版社总编办公室

邮 编：100036